SORT

Statistics and Operations Research Transactions

Volume
47

Number 1, January-June 2023

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Statistics and Operations Research Transactions

Volume 47, Number 1, January-June 2023

ISSN: 1696-2281
eISSN: 2013-8830

9 771696 228009 47001

www.idescat.cat/sort/

Aims

SORT (*Statistics and Operations Research Transactions*) -formerly *Qüestiió*- is an international journal launched in 2003, published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-edited by the Universitat Politècnica de Catalunya, Universitat de Barcelona, Universitat Autònoma de Barcelona, Universitat de Girona, Universitat Pompeu Fabra, Universitat de Lleida i Universitat Rovira i Virgili and the cooperation of the Spanish Section of the International Biometric Society, the Catalan Statistical Society and the Departament de Recerca i Universitats, of the Generalitat de Catalunya. SORT promotes the publication of original articles of a methodological or applied nature or motivated by an applied problem in statistics, operations research, official statistics or biometrics as well as book reviews. We encourage authors to include an example of a real data set in their manuscripts. SORT is an Open Access journal which does not charge publication fees.

SORT is indexed and abstracted in the *Science Citation Index Expanded* and in the *Journal Citation Reports* (Clarivate Analytics) from January 2008. The journal is also described in the *Encyclopedia of Statistical Sciences* and indexed as well by: *Current Index to Statistics*, *Índice Español de Ciencia y Tecnología*, *MathSci*, *Current Mathematical Publications and Mathematical Reviews*, and *Scopus*.

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa* (*Qüestiió*), published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestiió* have their origin in the *Cuadernos de Estadística Aplicada e Investigación Operativa*, published by the UPC till 1977.

Secretary and subscriptions to SORT

Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 557.30.76 - 93 557.30.00
Fax. +34 - 93 557.30.01
E-mail: sort@idescat.cat

Publisher: Institut d'Estadística de Catalunya (Idescat)

ISSN 1696-2281
eISSN: 2013-8830
DL B-46.085-1977
Key title: SORT
Numbering: 1 (december 1977)
www.idescat.cat/sort/

ISSN: 1696-2281
eISSN: 2013-8830
SORT 47 (1) January-June (2023)

SORT

Statistics and Operations Research Transactions

Coediting institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Universitat Pompeu Fabra
Universitat de Lleida
Universitat Rovira i Virgili
Institut d'Estadística de Catalunya

Supporting institutions

Spanish Region of the International Biometric Society
Societat Catalana d'Estadística
Departament de Recerca i Universitats

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Volume 47
Number 1
January-June 2023
ISSN: 1696-2281
eISSN: 2013-8830

Invited article

Articles

SORT 47 (1) January-June 2023, 3-80 DOI: 10.57645/20.8080.02.1

Transport systems analysis: models and data

Jaume Barceló[1]

Abstract

Rapid advancements in new technologies, especially information and communication technologies (ICT), have significantly increased the number of sensors that capture data, namely those embedded in mobile devices. This wealth of data has garnered particular interest in analyzing transport systems, with some researchers arguing that the data alone are sufficient enough to render transport models unnecessary. However, this paper takes a contrary position and holds that models and data are not mutually exclusive but rather depend upon each other. Transport models are built upon established families of optimization and simulation approaches, and their development aligns with the scientific principles of operations research, which involves acquiring knowledge to derive modeling hypotheses. We provide an overview of these modeling principles and their application to transport systems, presenting numerous models that vary according to study objectives and corresponding modeling hypotheses. The data required for building, calibrating, and validating selected models are discussed, along with examples of using data analytics techniques to collect and handle the data supplied by ICT applications. The paper concludes with some comments on current and future trends.

MSC: *90B20, 90B10, 90-08, 90-C25, 90B06.*

Keywords: *Optimization, Simulation, Data Analytics, Traffic assignment, Traffic Simulation.*

1. Introductory remarks on system modeling and data

Operations research (OR) has been a scientific discipline since its inception, as noted by Blackett (1948), Ackoff, Gupta and Minas (1965), and it adheres to the methodological principles of science. Barceló (2015) states that "systems are observed; observations consist of measurements that are data or, in other words, facts from which laws can be derived and which can be articulated in the body of theories". An epistemological chain in

[1] Department of Statistics and Operations Research. Universitat Politècnica de Catalunya-UPC Barcelona Tech. Professor Emeritus.

Received: May 2023

which, for Mario Bunge (1960), the theories are usually formalized in terms of mathematical models, representing parts of the reality, which can be *descriptive* or *predictive*.

Barceló further elaborates that while a descriptive model explains what happens, a description is often insufficient and it is typically necessary to delve into how things happen and, if possible, why. This gives rise to a need for predictive models. In other words, science usually aims to be predictive. To these epistemological foundations, especially in the case of OR, we could also add: "the recognition of uncertainty as an inherent component of observed reality, and the falsifiability principle of Popper (1972), a cornerstone of the scientific method: the theories or models are true until further empirical evidence proves them false".

In accordance with Heavens, Ward and Natalie (2013), we adopt the perspective that "a model formally organizes what we know, or we think we know, about a system to predict how it might behave in the present, future or past, as well as how it might respond to external influence" Furthermore, the first methodological step in studying a system involves data acquisition, which is carried out in accordance with the study objectives and available technologies for making observations, specifically through measurements. The underlying assumption of this step is that data contain information about the phenomenon under study, for which the data must be suitably processed and analyzed in order to find the necessary information. In other words, we can deduce how systems work by acquiring knowledge about them and translating this understanding into "laws" or modeling hypotheses, with the formal structure of these hypotheses defining the system model. In what follows, our present research on transport systems will employ mathematical formalism in terms of equations while also relying on implicit representations such as simulation models.

The model-building process consists of translating the modeling hypothesis, derived from the acquired knowledge, into appropriate formal terms that align with the study objectives. In the realm of operations research or similar perspectives adopted when analyzing transport systems, the models of analyzed systems serve a crucial purpose. Beyond acquiring knowledge about the studied system, their primary objective is to generate new insights by answering what-if questions regarding the system's response to external influences, such as transport policies that can impact its behavior. In other words, in the context addressed in this paper, a key objective of the modeling task is to utilize the model as the core component of a decision support system, aimed at facilitating optimal decision-making to enhance the system's response. Figure 1 provides a conceptual diagram for visualizing this methodological process.

Data have traditionally been scarce and costly, especially in the case of transport systems. However, this situation has dramatically changed with the advent and widespread adoption of new information and communication technologies (ICT), such that we are often overwhelmed by the deluge of data. This has led some to claim that the scientific method is no longer necessary. As Anderson (2008) famously argued, "Petabytes allow us to say: 'Correlation is enough.' [...] Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic

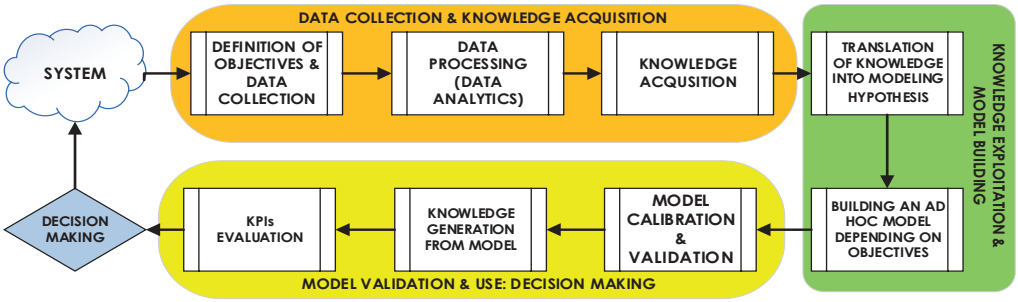

Figure 1. *Our methodological scheme for building and using models.*

explanation at all." Nevertheless, it is essential to heed the advice of the International Transport Forum (ITF, 2015), among many others. In the specific context of transport systems, ITF warns that the availability of massive and near real-time datasets can create the illusion of mirroring reality and tempt us to assume that data alone provide an accurate representation of reality. Consequently, there is a perceived notion that we can dispense with classic statistical tests regarding bias, validity, explanatory theories, and models. However, numerous examples demonstrate that data analytics have failed to provide long-term robust predictive results. Therefore, there is a need for algorithms that consistently detect patterns and mitigate bias in the data. These techniques are well suited to discovering less obvious or even hidden correlations in the initial data. We must bear in mind not only that correlation does not imply causation, but that they are completely different. Although correlated variables may reveal a possible causal relationship in the data, they do not explain which correlations are meaningful or predictive. Even if a correlation proves to be robust over a given period, data analytics alone cannot provide insights into factors that may cause the correlation to break down or lead to the emergence of new patterns.

The approach adopted in this paper assumes that both data and models are necessary. This notion is succinctly encapsulated by an anonymous quotation, which states that "Data without models are just numbers, but models without data are just stories."

The methodological approach summarized in Figure 1 highlights another relevant aspect: the objectives driving the analysis of the system.

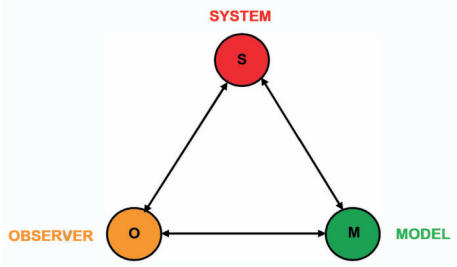

Figure 2. *The Minsky triad.*

These objectives serve as a guiding force for determining the relevant and necessary data to collect, as well as for constructing the system model. Thus, different models of the same system may exist, depending on the study objectives. Minsky's proposal in 1965 (Minsky, 1965) has come to be known as the Minsky triad of system, observer, and model (Figure 2), which we can summarize by saying that an object M is considered a model of a system S if it can provide valid answers to the questions posed by an observer O of that system.

Consequence: There is no such a thing as the unique model of a system.

This paper is structured as follows. Section 2 demonstrates how to apply the Minsky triad to transport systems, presenting a summary of the main alternative modeling approaches used in transport systems analysis. These approaches vary depending on the objectives and underlying modeling hypotheses. The section identifies the data requirements for each type of model and discusses the role played by data and the criteria for determining what makes a model useful. Section 3 provides some examples of the data needed to build these models, as well as how they are sourced through ICT applications and the necessary data analytics processes that render them usable. Section 4 summarizes the main conclusions and recommendations drawn from the paper, along with an overview of current and future trends in transport modeling.

2. Modeling transport systems

The pervasive penetration of the automobile as a private motorized transportation mode was driven by the economic interests of major manufacturers in the 1930s and gained momentum following the Second World War, resulting in a profound social and urban transformation of cities and metropolitan areas. The phenomena of urban sprawl described by Barceló (2019) emerged as a consequence of an unplanned and anarchic expansion due to a combination of factors: the relative affluence shifting from rural to city populations, lifestyle changes, and, particularly, advancements in individual motorized mobility. The latter factor led to a spatial separation of residential and working areas, made possible by the development of transportation systems that in turn led to the well-known consequences that we call traffic congestion.

The main consequence was a substantial demand for the expansion of road networks, thereby necessitating the development of appropriate tools for rational planning processes to assist decision-makers in determining which infrastructures to develop and how to meet the growing demand. Almost at the same time, the escalating traffic congestion sparked interest in understanding the dynamics of traffic flows and causes of congestion. The hope was that better comprehension of these factors would improve management policies and possibly alleviate the negative impacts associated with congestion.

Depending on the objectives, various models have been developed to analyze transport systems. This section provides a concise overview of the primary models that re-

flect a complementary understanding of transport systems, their associated modeling hypotheses, and their translation into mathematical models. Additionally, it provides insight into the data required to support these models, which are the following:

a. Models aimed at understanding how travelers use the existing road transport network to navigate from their origins to destinations in a given geographic area. These models support the long-term planning of transport infrastructures and are known as "static traffic assignment models."

b. Models for understanding the dynamics of traffic flows.

- b.1 Macroscopic models based on traffic flow theory, with an aggregated perspective of flow dynamics.
- b.2 Microscopic models that describe flow dynamics by considering the individual components constituting the flow.

c. Models explicitly accounting for the dynamics of traffic flows.

- c.1 Dynamic assignment models that explicitly account for time dependencies. These either analytically describe traffic flows or approximate their dynamics through simulation.
- c.2 Microscopic simulation models that capture the individual dynamics of vehicles within the traffic flow.

Note: The following focuses solely on the traffic flows of passenger cars. Public transport requires a similar but distinct modeling approach with specific features that are different from passenger cars. Including models for public transport in this paper would make it excessively long.

2.1. Static traffic assignment models

Transportation analysis typically revolves around understanding traffic patterns in a given geographic area, most frequently an urban or metropolitan area spanned by a transportation network. The goal is to gain insight into how the transport demand (i.e., the volume of trips in an area) uses the transport infrastructure under certain conditions. This transport demand is commonly defined in terms of an origin-to-destination (OD) matrix, X, whose entries (r,s) represent the number of trips from an origin r to a destination s. From a practical point of view, the study area is split into many transport analysis zones (TAZ) using well-established criteria that consider factors such as surface area and socioeconomic data obtained from various sources, for example, census tracts or population statistics (Ortúzar and Willumsen, 2011).

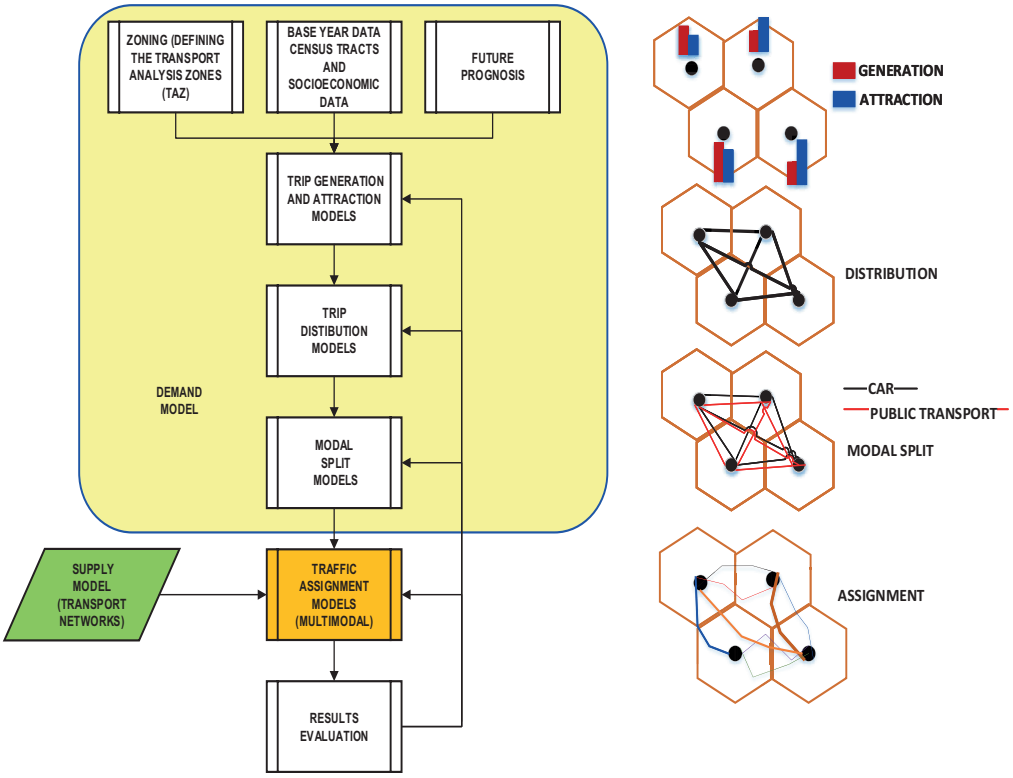

Figure 3. *Conceptual diagram of the four-step model.*

Figure 3 conceptually summarizes the logic diagram of the conventional planning process approach known as the *four-step model*, which has become a standard in transport modeling since the Second World War, due to a need to manage the consequences of post-war development and economic growth. It is based on the landmark study by Mitchell and Rapkin (1954), who applied analytical methods to establish a comprehensive framework. However, this model has deficiencies and limitations that have been widely discussed (McNally, 2000). The reasons for mentioning this model are manifold. First, it provides an overview of the modeling exercise described in Section 1, applied specifically to transport systems. Second, it helps identify the comprehensive path of intermediate models that led to the targeted model in this paper, which is the traffic assignment model. Finally, it serves to highlight the data requirements and the type of data needed to render the model operational.

In summary, assuming a homogeneous zonal splitting of the TAZ and using associated socioeconomic attributes such as population and economic activities, econometric models can be constructed to estimate the total number of trips O_r generated by a given origin r within a specific period (e.g., an average working day), considering all possible destinations within the other TAZ of that area. Similarly, equivalent attributes allow building econometric models to estimate the total number of trips D_s attracted by a given

destination s. Building such generation and attraction models for all origins and destinations is the first step. The second step consists of estimating the number of trips x_{rs} leaving origin r for destination s, which yields an estimation of the origin to destination (OD) matrix M of the total number of trips between all OD pairs (r,s) without distinguishing which transportation mode is being used. The third step involves the *modal split model*, whose purpose is to discriminate among available transportation modes, such as passenger cars, public transport bus, or public transport metro. The modal split model typically relies on discrete choice models that consider the perceived utilities of the travelers, as discussed by Ben-Akiva and Lerman (1985), Ben-Akiva and Bierlaire (1999). In the case of passenger cars, the fourth step is the traffic assignment step.

Traffic assignment refers to the process of allocating traffic demand, represented by an origin-destination matrix, onto the transportation network. This enables computing traffic flows on network links and offers insights into trip behavior and accessibility to activity locations. Figure 4 illustrates the four steps applied to the *first crown* of the Metropolitan Area of Barcelona, where first crown refers to the continuum of the 18 most populated municipalities including the city.

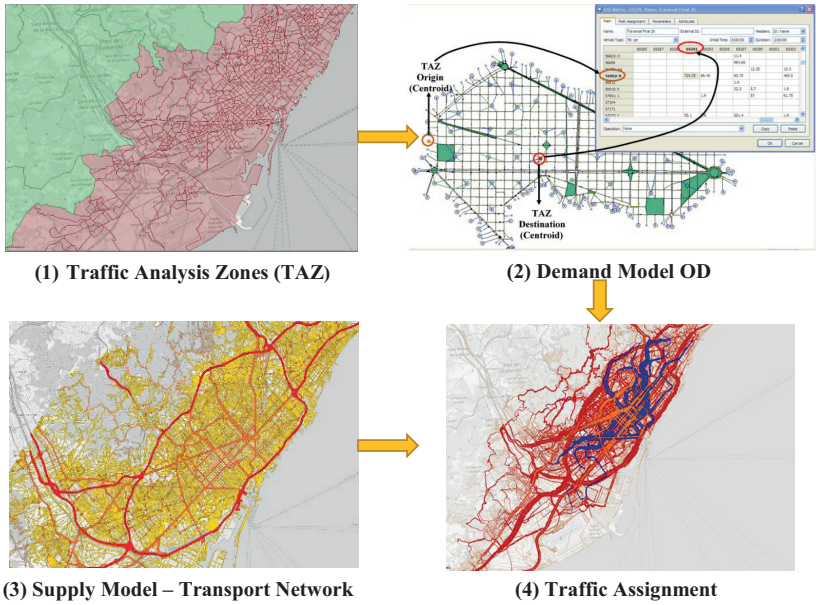

Figure 4. *(1) The partitioning of the first crown into TAZ. (2) The correspondence between the centroids representing an origin-destination pair for the subarea corresponding to the central business district of the city and the associated OD matrix. (3) Highlighting of the main arterials of the road network of the first crown. (4) The traffic flows resulting from a multimodal traffic assignment, with car traffic flows in red, bus flows in blue, and metro flows in orange. The thickness of the lines scales the intensity of the depicted flows.*

The underlying modeling hypothesis is that travelers move from origins to destinations in the network by selecting available routes based on behavioral choices governed

by certain rules. The characteristics of a traffic assignment procedure are determined by this key modeling hypothesis that is based on the concept of user equilibrium, which assumes that travelers try to minimize their individual travel times by choosing what they perceive to be the shortest routes under prevailing traffic conditions. This modeling hypothesis is formulated in terms of Wardrop's first principle (Wardrop, 1952):

> The journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route.

Traffic assignment models based on this principle are known as user equilibrium models, which differ from models whose objective is to optimize the total system travel time independently of individual preferences. For a more in-depth exploration that will not be considered here, see Sheffi (1985), Florian and Hearn (1995), and Patriksson (1994). Florian and Hearn (1995) demonstrated that when the path flows x_{rsp} from origin r to destination s along path p, with path costs tt_{rsp}, they must satisfy:

$$(tt_{rsp} - \theta_{rs})\ x_{rsp} = 0 \ \ \forall p \in K_{rs}\ \forall (r,s) \in I \tag{1}$$

$$tt_{rsp} - \theta_{rs} \geq 0 \ \ \forall p \in K_{rs}\ \forall (r,s) \in I \tag{2}$$

$$tt_{rsp},\ \theta_{rs},\ x_{rsp}\ \geq 0 \quad \forall p \in K_{rs}\ \forall (r,s) \in I \tag{3}$$

and the flow balancing equations:

$$\sum_{\forall p \in K_{rs}} x_{rsp} = X_{rs},\ \forall (r,s) \in I \tag{4}$$

where θ_{rs} image1 is the cost of the shortest path from r to s, K_{rs} is the set of all available paths from r to s, I is the set of all origin-destination pairs (r,s) in the network and X_{rs} is the demand (number of trips) from r to s. Then, these flows are in an equilibrium that satisfies Wardrop's principle. Effectively, if path p from origin r to destination s carries a flow $x_{rsp} > 0$, then the first equation is satisfied only if the path cost tt_{rsp} is equal to the minimum path cost θ_{rs} , that is, $tt_{rsp} - \theta_{rs} = 0$ for all paths from r to s, as required by Wardrop's principle. Reciprocally, if the path cost tt_{rsp} is greater than the minimum path cost θ_{rs}, that is $tt_{rsp} - \theta_{rs} > 0$, then satisfying the first equation requires that the flow on path p from r to s be zero, which is in other words an unused path according to Wardrop's principle.

Constraints (4) determine when a flow is feasible or not in terms of flow balance. If K_{rs} is the set of all paths for the (r,s) OD pair, then the sum of flows on paths for the (r,s) OD pair must be equal to the demand X_{rs}, which is the total number of trips for that OD pair. Applying some algebra (Florian and Hearn, 1995; Patriksson, 1994), the static traffic assignment model can be formulated in the space of the path feasible flows $\aleph$ in terms of the following system of variational inequalities:

$$T(X^*)(X - X^*) \geq 0,\ \forall X \in \aleph \tag{5}$$

where $T(X^*)$ is the vector of path flows, X^* is the optimal path flow, and $\aleph$ is the set of feasible path flows.

$$\aleph = \left\{ x_{rsp} \left| \sum_{\forall p \in K_{rs}} x_{rsp} = X_{rs},\ \forall (r,s) \in I\ ,\ x_{rsp} > 0 \right. \right\} \tag{6}$$

We assume that the road network is modeled in terms of a graph $G = (N, A)$ with a set of nodes N representing either intersections or dummy nodes associated with the transportation zones (usually referred to as centroids), and a set A of arcs used to model the infrastructure and the connections between centroids to the network. Thus, considering the relationships between path flows x_{rsp} and link flows $v_a,\ \forall a \in A$, we have:

$$v_a = \sum_{(r,s)\in I} \sum_{p \in K_{rs}} x_{rsp} \delta_{ap} \quad where \begin{cases} 1 & \text{if arc a belongs to path } p \text{ from } r \text{ to } s \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where δ_{ap} are the entries of the link-path incidence matrix. Assuming that the relationship between path costs tt_{rsp} and link costs $c_a(v)$, then:

$$tt_{rsp} = \sum_{a \in A} \delta_{ap} c_a(v) \tag{8}$$

where the link cost of each arc a $c_a(v),\ \forall a \in A$ is a function of the vector of feasible flows v in all arcs. Again, after some algebra (Florian and Hearn, 1995; Patriksson, 1994), the equivalent formulation of the model (5) in the space of link flows is:

$$C(v^*)[v - v^*] \geq 0,\ v \in V \tag{9}$$

$$V = \left\{ v : v_a = \sum_{(r,s)\in I} \sum_{p \in K_{rs}} x_{rsp} \delta_{ap}\ ,\ \sum_{\forall p \in K_{rs}} x_{rsp} = X_{rs},\ x_{rsp} \geq 0,\ \forall (r,s) \in I \right\} \tag{10}$$

where V is the set of feasible flows. This the Smith's (1979) variational inequality. It can be proven that there is no equivalent convex optimization problem unless the cost functions $c_a(v)$ are separable, meaning their Jacobian is symmetric (Florian and Hearn, 1995). The simpler separability condition holds when they depend only on the flow in the link:

$$c_a(v) = c_a(v_a),\ \forall a \in A \tag{11}$$

and demands X_{rs} are considered constant, independent of travel costs. Thus, the variational inequality formulation has the following equivalent convex optimization problem (Patriksson, 1994; Florian and Hearn 1995):

$$\begin{aligned} &\text{Min } C(v) = \sum_{a \in A} \int_0^{v_a} c_a(x) dx \\ &s.t. \sum_{\forall p \in K_{rs}} x_{rsp} = X_{rs},\ \forall (r,s) \in I \\ &(x_{rsp} \geq 0\ \forall (r,s) \in I,\ p \in K_{rs} \end{aligned} \tag{12}$$

and the definitional constraint of v_a (7).

Assuming that the separability conditions hold and that therefore the link cost functions of each link a depend only on the volume v_a in that link, a critical aspect of the model-building process is determining the specific form of the functions $c_a(v_a)$. This assumption is also relevant for numerically solving the model's algorithms. The link cost functions $(c_a(v_a)$, also known as *volume delay functions* (VDF), quantify the variation in travel time within a link a according to the traffic volume v_a. This reflects the dependencies between a link's travel time and its traffic flow, although the link cost functions can also include additional factors indicating the cost (or impedance) of using a link, such as toll fares in urban pricing systems. In such cases, the functions represent the generalized costs of using the link, and they can be interpreted as indicators of the level of service provided by the link. Empirical studies suggest that these costs play a crucial role when users decide which routes to take.

The first form proposed for the VDF functions was that introduced by the Bureau of Public Roads in 1964 and has since become widely known as BPR functions, which are still widely used in the current transportation modeling practice. They take the analytical form:

$$c_a(v_a) = t_0^a \left[1 + \alpha_a \left(\frac{v_a}{\kappa_a}\right)^{\beta_a}\right] \tag{13}$$

where t_0^a represents the minimum time to traverse link a, which traffic engineers call the free flow time, which represents the time it takes for a to travel freely through the link without competing with other vehicles for the available capacity. κ_a represent the capacity or maximum flow of link a; and α_a and β_a are link-specific parameters that must be calibrated.

Although BPR functions are appealing for their simplicity and relative ease of calibration, they nevertheless exhibit anomalous behavior in certain circumstances. For example, when the β_a is large, it provides abnormal values for those links with $\frac{v_a}{\kappa_a} > 1$ in the first iterations of the assignment algorithms. This delays the convergence due to the anomalous overload of those links. Additionally, for links with traffic loads well below their capacity, the link travel time remains nearly equivalent to the free flow time regardless of the actual flow. These shortcomings prompted the search for more sophisticated functions that overcome these drawbacks. One well-known family of functions, widely used in practice, is the family of conical functions proposed by Spiess (1990):

$$c_a(v_a) = t_0^a \left[2 - \beta_a - \alpha_a\left(1 - \frac{v_a}{\kappa_a}\right) + \sqrt{\alpha_a^2\left(1 - \frac{v_a}{\kappa_a}\right)^2 + \beta_a^2}\right],\ \beta_a = \frac{2\alpha_a - 1}{2\alpha_a - 2},\ \alpha_a > 1 \tag{14}$$

Akcelik (1991) proposes an alternative family of VDF that explicitly accounts for the delays incurred by traffic lights at signalized intersections:

$$c_a(v_a) = t_0^a \left\{1 + 0.25\frac{T}{t_0^a}\left[\frac{v_a}{\kappa_a} - 1 + \sqrt{\left(\frac{v_a}{\kappa_a} - 1\right)^2 + \frac{8J_a}{\kappa_a T}.\frac{v_a}{\kappa_a}}\right]\right\} \tag{15}$$

where T is the duration of the calibration interval, and $J_a$is the delay factor for link a, as defined by:

$$J_a = \frac{d}{1 + \kappa_a T}$$

where d is a delay parameter whose value is $d = 1$ for exponential arrivals, and $d = 0.5$ for uniform arrivals. In all cases, the proposed VDF depends on parameters that must be calibrated, which requires observational data (Petrik, Moura and Abreu e Silva, 2014; Bessa, de Magalhães and Santos, 2021).

One common characteristic of all the proposed VDFs is that they are convex. Consequently, the symmetric static traffic assignment problem in user equilibrium, defined by (11), is a non-linear convex optimization problem. The properties of the convexity can provide benefits for the algorithmic approaches to solving the problem numerically.

Although the traffic assignment problem is a specific case of the non-linear multicommodity network flows problem and can be solved by any of the methods used for solving such problems, more efficient algorithms have been developed (LeBlanc, Morlok and Pierskalla, 1975; Florian and Nguyen, 1976) by adapting the linear approximation method of Frank and Wolfe (1956). Other efficient algorithms (Hearn, Lawphonpanich and Ventura, 1987; Lawphongpanich and Hearn, 1984) are based on the restricted simplicial approach, which exploits the properties of the convex polyhedron of feasible solutions defined by the constraints outlined in equation (10). Additionally, the parallel tangents method (PARTAN) introduced by Florian, Guelat and Spiess (1987) has proven to be effective. This version of the model is preferred by the main professional software platforms for transport planning because of its algorithmic ability to computationally deal efficiently with large road networks, particularly those found in large metropolitan areas. Consequently, research efforts have focused on improving these algorithms for solving the problem. Other contributions include those analytical approaches inspired by the gradient projection methods of Rosen (1960), which exploit the properties of the polytope defined by constraints (4) and combine them with the efficient shortest paths algorithms. Notable examples of these contributions can be found in Bar-Gera (2002), Dial (2006), and Florian and Constantin (2009).

Criticism of the symmetric traffic assignment problems arises from their inability to properly answer more demanding modeling requirements, due to the inaccuracies induced by the oversimplified separability assumptions in the VDF. This may happen, for instance, when dealing explicitly with delays at unsignalized intersections or when the generalized costs in multiclass planning models depend on vehicle class interactions that induce asymmetries. To address these issues, is necessary to develop alternative formulations that account for the asymmetries in either the space of the path flows (as in equations (5) and (6)) or in the space of the link flows (as in equations (9) and (10)). These models, known as asymmetric traffic assignment (ATA) in user equilibrium, are more appropriate for tackling the problem.

Equations (5), (6), (9), and (10) are typical variational inequality (VI) formulations in finite dimension spaces that correspond to an equilibrium principle. In essence, they

can be described as follows (Codina, Ibáñez and Barceló, 2015): Given a closed convex set $X \in \Re^n$ of candidate solutions in which a continuous function $F(\cdot) : X \to \Re^n$ is defined, we look for a special point x such that the projection $x - \alpha F(x)$, for any $\alpha > 0$, onto X results in the point x itself. In other words, if $P_X[\cdot]$ is the projection operator on X defined by

$$P_X[z] = \text{argmin}\left\{ \left(\|z-x\|_2^2 \middle| x \in X \right) \right\}$$

then a solution x of the VI verifies the fixed-point relationship

$$x = P_X\left[x - F(x)\right]$$

which is equivalent to stating that the solutions to VI problems satisfies the condition

$$F(x)^T(y-x) \geq 0,\ \forall y \in X$$

Rewriting (7) in vector form, we have $v = \Delta X$, and the VI (9) can thus be rewritten (Florian and Hearn, 1995) as:

$$C\left(\Delta X^*\right)\left[\Delta X - \Delta X^*\right] \geq 0,\ X \in \aleph$$

That is

$$\Delta^T C\left(\Delta X^*\right)\left(X - X^*\right) \geq 0,\ X \in \aleph \tag{16}$$

which can be solved by the projection algorithm:

$$X^{l+1} = P_{Q,\aleph}\left[X^l - \rho Q^{-1}\Delta^T C\left(\Delta^T X^l\right)\right]$$

and is equivalent to the convex optimization problem:

$$\text{Min}_{X\in\aleph}\left(X - X^l\right)\overline{C}(X) + \frac{1}{2\rho}\left(X - X^l\right)Q\left(X - X^l\right) \tag{17}$$

With $\overline{C}(X) = \Delta^T C\left(\Delta X^l\right)$, and Q a block-diagonal symmetric definite positive matrix (Codina et al., 2015), we have:

$$Q = \text{diag}\left[\ldots Q^{rs}\ldots;\ (r,s) \in I\right]$$

With each block Q^{rs} corresponding to an OD pair (r,s):

$$Q^{rs} = \text{diag}\left(\ldots q^{rsp}\ldots; p \in K_{rs}\right)$$

Then, (17) can be decomposed into the sequence of quadratic optimization problems, one for each OD pair:

$$\text{Min}\sum_{p\in K_{rs}}\left(x_{rsp} - x_{rsp}^{l-1}\right)\overline{C}_{rsp}\left(x_{rsp}^{l-1}\right) + \frac{1}{2\alpha_p^{-1}}\left(x_{rsp} - x_{rsp}^{l-1}\right)^2 \tag{18}$$

$$s.t.\ \sum_{p\in K_{rs}} x_{rsp} = X_{rs},\ x_{rsp} \geq 0$$

With $\alpha_p = \frac{\delta}{q^{rsp}}$ and δ a scaling parameter. Algorithms for numerically solving the subproblems proceed iteratively by generating new paths at each iteration while working in the sub-polytope of polytope $\aleph$ of the already identified paths and path flows. Codina et al. (2015) discuss the pros and cons of the alternative formulations of ATA in terms of VI, as well as the various algorithmic approaches for numerically solving them.

2.2. Dynamic traffic assignment models

With the emergence of intelligent transport systems (ITS), advanced traffic management systems (ATMS), and advanced traffic information systems (ATIS) as the most relevant ITS applications, planners need dynamic models that capture the time-dependent nature of changing traffic flows and traffic demand. The dynamic traffic assignment (DTA) problem can thus be considered an extension of the traffic assignment problem described above. DTA can determine time-varying links and path flows, and thus the temporal and spatial evolution of traffic flow patterns in the network (Mahmassani, 2001). The problem can be formulated as a dynamic user equilibrium problem built on the dynamic version of Wardrop's principle (Friesz et al., 1993; Smith, 1993; Ran and Boyce, 1996):

> If, for each OD pair at each instant of time, the actual travel times experienced by travelers departing at the same time are equal and minimal, the dynamic traffic flow over the network is in a state of travel-time-based dynamic user equilibrium (DUE).

Similarly to the translation of the static Wardrop's principle in into the variational inequalities (5) and (6), the DUE approach can also be implemented by solving the following mathematical model:

$$\begin{aligned} [tt_{rsp}(t) - \theta_{rs}(t)]\, x_{rsp}(t) = 0, \qquad & \forall p \in K_{rs}(t),\ \forall (r,s) \in I,\ t \in [0,T] \\ tt_{rsp}(t) - \theta_{rs}(t) \geq 0, \qquad & \forall p \in K_{rs}(t),\ \forall (r,s) \in I,\ t \in [0,T] \\ tt_{rsp}(t),\ \theta_{rs}(t),\ x_{rsp}(t) > 0, \qquad & \forall p \in K_{rs}(t),\ \forall (r,s) \in I,\ t \in [0,T] \end{aligned} \tag{19}$$

and the flow balancing equations

$$\sum_{\forall p \in K_{rs}(t)} x_{rsp}(t) = X_{rs}(t),\ \forall (r,s) \in I,\ t \in [0,T] \tag{20}$$

where, as before, $x_{rsp}(t)$ is the flow on path p from r to s, departing origin r at time interval t; $tt_{rsp}(t)$ is the actual path cost from r to s on route p at time interval t; θ_{rs} is the cost of the shortest path from r to s, departing from origin r at time interval t; $K_{rs}(t)$ is the set of all available paths from r to s at time interval t; I is the set of all origin-destination pairs (r,s) in the network, and $X_{rs}(t)$ is the demand (number of trips) from r to s, departing r at time interval t.

This is equivalent to solving a finite-dimensional variational inequality problem for finding a vector of path flows $\boldsymbol{x}^*$ and a vector of path travels times $\boldsymbol{\tau}$, such that:

$$\left[\mathbf{x} - \mathbf{x}^*\right]^{\mathsf{T}} \boldsymbol{\tau} \geq 0,\ \forall \mathbf{x} \in \aleph \tag{21}$$

where $\aleph$ is the set of feasible flows defined by:

$$\aleph = \left\{ x_{rsp}(t) \left| \sum_{\forall p \in K_{rs}(t)} x_{rsp}(t) = X_{rs}(t),\ \forall (r,s) \in I,\ t \in [0,T]\ ,\ x_{rsp}(t) > 0 \right. \right\} \tag{22}$$

Wu et al. (1991), Wu, Chen and Florian (1998a), and Wu et al. (1998b), prove that this is equivalent to solving the discretized variational inequality:

$$\sum_{t \in [0,T]} \sum_{p \in \Re} tt_{rsp}(t) \left[x_{rsp}(t) - x^*_{rsp}(t) \right] \geq 0 \tag{23}$$

where $\Re = \bigcup_{(r,s) \in I} K_{rs}$ is the set of all available paths. This can be solved numerically with ad hoc projection algorithms, which are described in these references.

2.3. Models based on traffic flow theory

The approaches described so far are treat a trip as an analytical unit associated with an individual while also assuming that the two are separate and independent. Although path and link flows result from aggregating these trips, the proposed models implicitly ignore their nature.

However, an alternative hydrodynamic perspective views the temporal propagation of traffic flows as analogous to a fluid flowing through the network. This alternative modeling perspective aligns with Minsky's statement that a system can be modeled in different ways according to various approaches and the modeler's objectives.

This hydrodynamic analogy can be approached in two ways. One takes an aggregate perspective that focuses on the overall state of the fluid using aggregate macroscopic variables for density, volume, and speed. The other delves into the dynamics of the fluid by taking a fully disaggregated point of view that aims to describe the fluid process in terms of its constituent individual particle dynamics (the vehicles). Complete descriptions of these approaches can be found in Barceló (2010) or in Chapters 7 (Hydrodynamic and Kinematic Models of Traffic) and 6 (Car Following and Acceleration Noise) in the monograph *Traffic Flow Theory*, by Gerlough and Huber (1975).

Two independent papers published almost simultaneously (Lighthill and Whitham, 1955; Richards, 1956), introduced the fundamental principles of the hydrodynamic analogy for modeling traffic flows. This approach, known as the Lighthill–Whitham–Richards (LWR) model, considers a motorway with two counting stations, CS-1 and CS-2, as depicted in Figure 5, separated by a distance ΔX.

Let us first assume that traffic flows in the direction of the arrow, and that N_1 and N_2 represent the number of vehicles counted during the time interval Δt at CS-1 and CS-2, respectively. Then, if $N_1 > N_2$, there is an accumulation $N_2 - N_1 = \Delta N$ of cars between the two counting stations during time interval Δt, as there are no sources or sinks of cars in that segment.

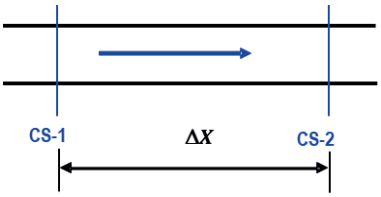

Figure 5.

The traffic flows are defined as volumes q_1 and q_2 passing through counting stations CS-1 and CS-2 during time interval Δt, using:

$$\frac{N_1}{\Delta t} = q_1 \;\; and \;\; \frac{N_2}{\Delta t} = q_2 \tag{24}$$

Assuming that the density k (the number of cars per unit distance) is homogeneous in the space between the counting stations during the considered time interval, its variation over the distance between them is given by:

$$\Delta k = \frac{-(N_2 - N_1)}{\Delta x} = \frac{-\Delta N}{\Delta x}$$

and thus

$$\Delta k \cdot \Delta x = -\Delta N \tag{25}$$

Similarly, the flow variation $\Delta q = q_2 - q_1$ during time interval Δt while taking into account (24) will lead to

$$\Delta q \cdot \Delta t = \Delta N \tag{26}$$

Assuming the modeling hypothesis of the conservation of cars, that is the flow conservation, it follows from (25) and (26) that

$$\frac{\Delta q}{\Delta x} + \frac{\Delta k}{\Delta t} = 0 \tag{27}$$

Since the medium can be considered as a continuum, then we can take infinitesimal intervals to express equation (27) as:

$$\frac{\partial q}{\partial x} + \frac{\partial k}{\partial t} = 0 \tag{28}$$

which is the continuity equation for a fluid.

A simple model for a highway stretch splits it into contiguous sections and models the dynamics of the traffic flow in each one using equation (28). The resulting discretized model, shown in Figure 6, discretizes the traffic state variables–density k and flow q–in space and time, such that k_i^j and q_i^j represent, respectively, vehicle density per kilometer and vehicle flow per hour for cell i at instant j. The upper and lower rows respectively describe the states of the discretized cells at two consecutive instants, j and $j+1$. The

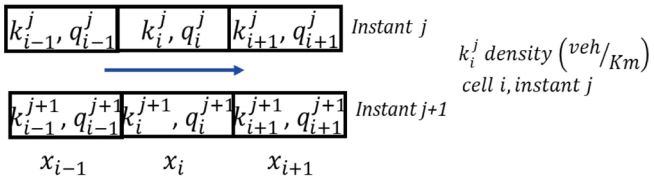

Figure 6. *Discretization of model (28).*

flow balance in cell i at time $j+1$, assuming that the inflow is equal to the outflow, can be expressed as:

$$k_i^{j+1}\Delta x = k_i^j \Delta x + q_{i-1}^j \Delta t - q_i^j \Delta t \tag{29}$$

This simpler model can be extended to highway sections with on- and off-ramps, like the one in Figure 7. These ramps serve as sources of incoming and outgoing traffic flows that are characterized by, respectively, rates $r(t)$ and $s(t)$. Thus, the continuity equation (28) (Michalopoulos, Beskos and Lin, 1984) becomes:

$$\frac{\partial q}{\partial x} + \frac{\partial k}{\partial t} = r(t) - s(t) = g(x,t) \tag{30}$$

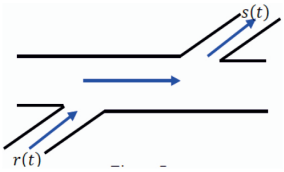

Figure 7.

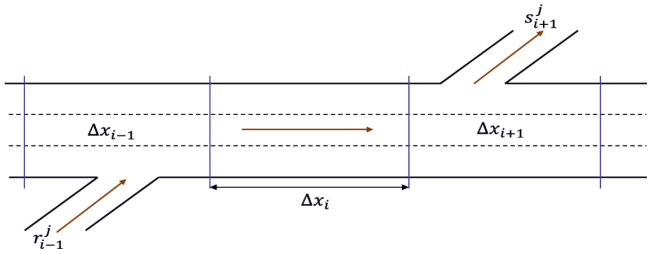

Figure 8. *Discretization to numerically solve the continuum motorway model.*

Defining $g(x,t)$ as the generation/dissipation function that balances the entry/exit flows, Michalopoulos (1984) generalized this model for motorways, as shown in Figure 8.

The time-space discretization of the model can be represented by the following set of difference equations:

$$k_i^{j+1} = \frac{1}{2}\left(k_{i+1}^j + k_{i-1}^j\right) - \frac{\Delta t}{2\Delta x}\left(q_{i+1}^j - q_{i-1}^j\right) + \frac{\Delta t}{2\Delta x}\left(g_{i+1}^j - g_{i-1}^j\right), \forall i \in I \tag{31}$$

where I is the set of cells.

This discretized form of the conservation equation is completed by an equation relating flows and densities, usually taking the form:

$$q_i^j = Q_i\left(k_i^j\right) \tag{32}$$

The set of equations (31) and (32) is usually known as the first-order macroscopic traffic flow model. For stability reasons, the time-space discretization in these models must satisfy the condition $\Delta x_i > u_f \Delta t_i$.

The interest in investigating the relationships between flows and densities arise simultaneously with the development of early traffic flow theories. These relationships came to be better understood by both measurement-based empirical evidence and theoretical analysis, such as Edie's (1963) seminal work. Let us assume a time-space diagram like the one depicted in Figure 9, where the traffic flows in spatially and homogenous conditions. The blue lines represent the vehicle trajectories in these conditions.

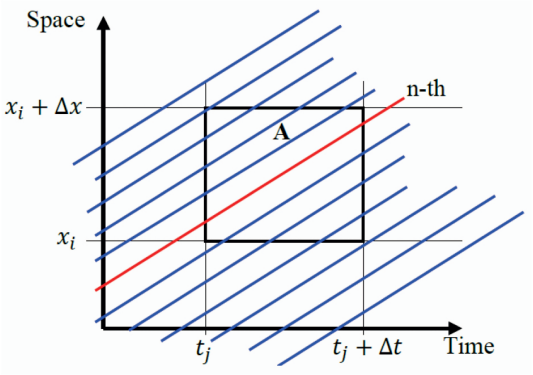

Figure 9. *Vehicle trajectories in the time-space diagram.*

Let us assume that one counting station is located at position x_i and the next counting station downstream is located at a distance Δx, corresponding to the location $x_i + \Delta x$. Let us also assume that Δt is the detection time resolution. Then, consecutive detectors and the detection time resolution define a discrete time-space Eulerian region A. Each vehicle (the n-th vehicle) follows a trajectory highlighted in red along a distance d_n during a time τ_n within this region. If $|A|$ is the surface of the region (in kilometers x hours), $d(A)$ is the total distance traveled by all vehicles crossing the region, and $t(A)$ represents the total time spent in the region by vehicles crossing it (in vehicles $\times$ hour). Defining the following quantities:

$$\text{Traffic Flow } \left(\frac{\text{vehicles}}{\text{hour}}\right) \qquad q(A) = \frac{d(A)}{|A|}$$

$$\text{Density } \left(\frac{\text{vehicles}}{\text{kilometer}}\right) \qquad k(A) = \frac{t(A)}{|A|}$$

$$\text{Mean spatial speed } \left(\frac{\text{kilometers}}{\text{hour}}\right) \qquad u(A) = \frac{d(A)}{t(A)}$$

the main relationship that emerges is $q(A) = u(A) \cdot k(A)$ or in the general form proposed by Daganzo (1997):

$$\mathbf{q} = \mathbf{uk} \tag{33}$$

which is known as the "fundamental traffic diagram". In the assumed stationary and homogeneous traffic conditions, it is further assumed that the mean spatial speed or equilibrium speed, denoted as $u_e(k)$ (Kotsialos and Papageorgiou, 2001) is a decreasing function of the density, which, combined with (33), leads to an equilibrium flow $q_e(k)$ that can be defined as:

$$q_e(k) = ku_e(k) \tag{34}$$

This corresponds to the steady state flow and homogeneous conditions. The fundamental diagram is a function that exhibits zeros at two extreme values of the density: $k = 0$ and $k = k_{\text{jam}}$, where the latter represents the jam density. This function has a unique maximum at the critical density $k = k_{cr}$, which corresponds to the maximum flow (capacity) q_{Max}.

Many mathematical expressions have been proposed for the function $u_e(k)$, based on either theoretical or empirical grounds. One of the most general is defined by the family of functions:

$$u_e(k) = u_f \left[1 - \left(\frac{k}{k_{\text{jam}}}\right)^{\alpha}\right]^{\beta} \tag{35}$$

where $u_f$is the free flow speed and α and β are parameters that must be calibrated. In this case, the complementary equations (32) of the first-order model for each section become:

$$q_i^j = k_i^j u_i^j \rightarrow \; u_i^j = u_f \left[1 - \left(\frac{k_i^j}{k_{\text{jam}}}\right)^{\alpha}\right]^{\beta} \tag{36}$$

Macroscopic models of traffic flows also provide additional examples of alternative models for the same system. First-order models can be extended to what is known as second-order models. These models, proposed by Papageorgiou and Kostsialos (2001) and TRB (2001), consider the mean speed as an independent variable and, therefore, require an extra equation for the speed dynamics, known as the momentum or speed equation. This introduces an additional modeling assumption that drivers respond to downstream traffic conditions with a corresponding reaction time, τ. Thus, the mean speed adjusts to the traffic density according to:

$$u(x,\, t+\tau) = u_e\left[k(x+\Delta x, t), t\right] \tag{37}$$

where Δx is the space increment. The expression (37) can be expanded using a Taylor series and, after rearranging the terms appropriately, we obtain:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \frac{1}{\tau}\left[u_e(k) - u - \frac{\upsilon}{k}\frac{\partial k}{\partial x}\right] \tag{38}$$

where υ is a parameter whose value, according to Payne (1971), should be $\upsilon = -0.5\frac{\partial u_e}{\partial k}$. This parameter υ is usually interpreted in terms of viscosity. To solve equation (38) numerically, as with other models, it can be discretized in both time and space:

$$u_i^{j+1} = u_i^j + \frac{\Delta t}{\tau}\left\{ u_e\left(k_i^j\right) - u_i^j + \frac{\Delta t}{\Delta x} u_i^j \left[u_{i-1}^j - u_i^j\right] - \frac{\upsilon \Delta t \left[k_{i+1}^j - k_i^j\right]}{\tau \Delta x \left[k_i^j + \kappa\right]} \right\} \quad (39)$$

Papageorgiou, Blosseville and Hadj-Salem (1990) propose including the term:

$$\frac{\delta \Delta t}{\Delta x}\left[\frac{r_i^j u_i^j}{k_i^j + \kappa}\right]$$

To model the impact of the entering on-ramp flows, δ and κ are additional model parameters whose values are estimated in the calibration process. The numerical solution to the second-order model is determined by equations (31), (36), and (39).

A variety of numerical methods for solving these macroscopic traffic flow models were developed between the 1970s and 1990s (Payne, 1971; Stock et al., 1973; Payne, 1979; Michalopoulos et al., 1984; Michalopoulos, 1984; Michalopoulos, Yi and Lyrintzis, 1993; Papageorgiou, Blosseville and Hadj-Salem, 1989; Papageorgiou, 1998).

2.4. Microscopic approaches

Continuing with the modeling exercise, we note here that transport and traffic systems offer fascinating examples of systems that can be modeled in various ways, depending on the objectives and underlying hypotheses. Let us now consider the traffic streams from a different perspective: a fully disaggregated standpoint aiming to explain the flow of the fluid in terms of its constituent individual particle dynamics. In the case of traffic flows, the individual particles are cars. One of the first attempts to describe flow dynamics in this way was the analysis conducted by Pipes (1953), who considered a line of traffic composed of n vehicles, as depicted in Figure 10, where L_k denotes the length of the k-th car and x_k its position. The modeling hypothesis is that the movement of several vehicles is controlled by a **law of separation**, which mandates that each vehicle must maintain a prescribed **following distance** from the preceding vehicle. This prescribed following distance, denoted as b, is proportional to the velocity of the following vehicle plus the minimum distance of separation when the vehicles are at rest.

Under this modeling hypothesis, the relationships between the positions of the consecutive vehicles, specifically the leader (vehicle k) and the follower (vehicle $k+1$) at time t are given by:

$$x_k(t) = x_{k+1}(t) + [b + T v_{k+1}(t)] + L_k \quad (40)$$

where T is a time constant whose value is such that $Tv_{k+1}(t)$ satisfies the law of separation. From (40), the first derivative with respect to time provides the relationships between the speeds:

$$\dot{x}_k(t) = \dot{x}_{k+1}(t) + T\dot{v}_{k+1}(t) \rightarrow v_k(t) = v_{k+1}(t) + T\dot{v}_{k+1}(t) \quad (41)$$

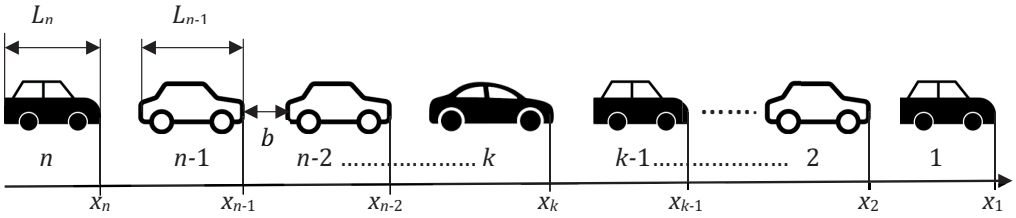

Figure 10. *Pipes' postulated line of traffic with n vehicles.*

which are the dynamical equations of the vehicle systems. Applying Laplace transform $\mathcal{L}$ and making $\mathcal{L}v_k(t) = V_k(p)$, this becomes the set of algebraic equations:

$$(Tp+1)V_{k+1}(p) = V_k(p) + Tpv_k(0) \tag{42}$$

If the velocity of the leader vehicle changes over time, then according to $v_1(t) = F(t)$, the velocity of vehicle $k+1$ is given by:

$$v_{k+1}(t) = \left[\frac{T^{-k}}{(k-1)!}\right] \int_0^t u^{k-1} \exp\left(-\frac{u}{T}\right) F(t-u)du \tag{43}$$

These dynamical equations are based solely on physical considerations without taking into account that cars are driven by humans and are thus subject to certain conditions. This model tries to replicate the dynamic behavior of a vehicle stream, where vehicles follow one another and adjust their acceleration or deceleration in order to maintain a prescribed separation distance for safety reasons. However, no assumptions are made regarding how drivers and vehicles achieve this objective. Newell (1961) adopts a different perspective that will later be incorporated into a more generalized approach, in which empirical evidence leads to assuming a nonlinear relationship between a car's velocity at time t and the spatial headway a short time before (i.e., at time $t-\Delta$). Additionally, families of velocity–headway relationships align well with experimental data for steady flows. Building upon these observations, Chandler, Herman and Montroll (1958), Herman, Montroll and Potts (1959), and Gazis, Herman and Potts (1959) used empirical observations in the well-known General Motors experiments to lay the foundations of the car-following theory to model the dynamics of both vehicle and driver, which can be summarized as follows.

- The study of traffic is a combination of experimental and observational science.
- It adopts the perspective of the theory of servomechanisms, a branch of applied mathematics.
- It aims to clarify the role and interaction of the three components of the traffic system:
  - Road topology, which includes the number of lanes, nature of intersections, signals, warning signs, and other related factors.

– Vehicle characteristics, encompassing speed, acceleration and deceleration qualities, and other relevant attributes.
– Driver behavior, such as the range of perception and the lags between perception and response.
– In order to develop a theory of stable car-following, a traffic element should be considered as a servomechanism. The conceptual conditions for such a theory are displayed in Figure 11 (from Rothery, 2001).

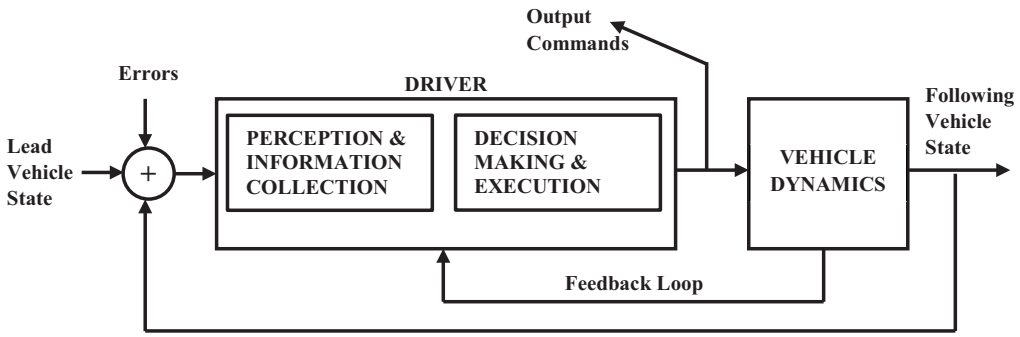

Figure 11. *Block diagram of the linear car-following model (Rothery, 2001).*

The main modeling hypotheses can be summarized as follows.

- Modeling assumption: Each driver reacts in a specific manner to stimuli from the preceding car(s), as stated by Gazis, Herman and Rothery (1961).
- Discrete cars move in continuous space and time.
- The laws of motion for each vehicle model driver behavior use differential-difference equations, which express the idea that each driver responds to a given stimulus according to a relationship defined by the following expression:

$$\text{RESPONSE} = \text{SENSITIVITY} \times \text{STIMULUS}$$

- The stimulus is a function of car positions, their time derivatives, and possibly other parameters.
- The response corresponds to the vehicle's acceleration or deceleration, as the driver has direct control through the gas and brake pedals.

Let us consider a situation like the one depicted in Figure 12, where various variables are defined as follows: x_n identifies the position of the *n-th* vehicle, $\dot{x}_n(t) = v_n(t)$ denotes its velocity at time t, and l_n its length. Additionally, $s_n(t) = x_{n-1}(t) - x_n(t)$ represents the spacing or space headway between the leader $n-1$ 1 and the follower n at time t,

which is also called the effective length of vehicle n (vehicle length + safety distance). The relative velocity at time t between a leader-follower pair, i.e. n (leader) and $n+1$ (follower), can be calculated as $\Delta v_{n+1}(t) = v_n(t) - v_{n+1}(t) = \dot{x}_n(t) - \dot{x}_{n+1}(t) = \dot{s}_{n+1}(t)$. The acceleration of vehicle n at time t is denoted by $a_n(t) = \ddot{x}_n(t)$.

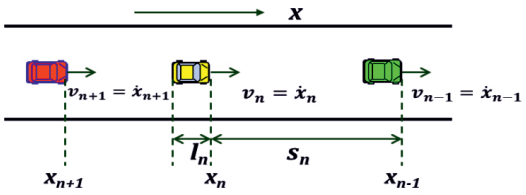

Figure 12. *Leader-Follower relationships.*

The car-following theory is based on the empirical observation that a strong correlation exists between a driver's response and the relative speed between their vehicle and the one ahead. The main modeling hypothesis is that the stimulus for a driver is the relative speed between each leader-follower pair, and the resulting mathematical model is a stimulus-response equation that describes the motion of the $(n+1)$-th car following the n-th car. In other words, the driver of the $(n+1)$-th vehicle observes variations in $v_n(t)$ or $s_n(t)$ and then accelerates or brakes to keep from lagging behind or getting too close to the leader. The mathematical model translating this hypothesis into formal terms is:

$$\frac{dv_{n+1}(t)}{dt} = F\left\{v_{n+1}(t); f_1\left[v_n(t) - v_{n+1}(t)\right]; s_{n+1}(t)\right\} \tag{44}$$

$$\frac{d^2x_{n+1}(t)}{dt^2} = \lambda\left[\frac{dx_n(t)}{dt} - \frac{dx_{n+1}(t)}{dt}\right]_{t-\tau}$$

Or, equivalently:

$$\ddot{x}_{n+1}(t+\tau) = \lambda\left[\dot{x}_n(t) - \dot{x}_{n+1}(t)\right] \tag{45}$$

This is a law for acceleration in a linear system, where λ represents the sensitivity of the control mechanism and τ is the time-lag of the driver-car system, which can be interpreted as the driver's reaction time. Integrating the model allows us to obtain the velocity of the n + 1-*th* vehicle, which represents the velocity of the traffic stream. Assuming that s = x_n -x_{n+1} is the average spacing (s = *1/k*) and when velocity u = *0*, then spacing s_{jam} = jam spacing $= 1/k_{\text{jam}}$, and k_{jam} = jam density. The integration of equation (45) yields:

$$u = \lambda\left[\frac{1}{k} - \frac{1}{k_{\text{jam}}}\right] \quad \textit{and } q = uk = \lambda\left[1 - \frac{k}{k_{\text{jam}}}\right] \tag{46}$$

which is the Greenshields, Gerlough, and Huber (1975) fundamental diagram of traffic. A detailed analysis of the model (44) reveals that the model is inconsistent with real data measurements. However, equation (46) indicates that it is conceptually consistent with the postulates of traffic flow theory.

Looking to improve the car-following model, Gazis et al. (1959) assume that the sensitivity λ varies with the distance between vehicles, as $\frac{\lambda}{s_{n+1}}$. In other words, the model takes into account that drivers' reactions will be quicker for denser traffic. Consequently, the updated model is as follows:

$$\ddot{x}_{n+1}(t+\tau) = \lambda \left[\frac{\dot{x}_n(t) - \dot{x}_{n+1}(t)}{x_n(t) - x_{n+1}(t)} \right] \tag{47}$$

which also exhibits inaccuracies. However, by integrating the equation once again to obtain the velocity of the $n+1$-*th* vehicle (velocity of the traffic stream), and assuming that the stream velocity u is $u = 0$ for $k = k_{\text{jam}}$, the integration yields:

$$u = \lambda ln \left(\frac{k_{\text{jam}}}{k} \right) \tag{48}$$

That is Greenberg's fundamental diagram. Despite its inaccuracies, the car-following model (47) remains consistent with the fundamental principles of traffic flow theory. In seeking to improve car-following models, Edie (1963) proposed a modified model in which the sensitivity λ will depend on the square of the spacing and the current speed, resulting in:

$$\ddot{x}_{n+1}(t+\tau) = \lambda \frac{\dot{x}_{n+1}(t)}{\left[x_n(t) - x_{n+1}(t)\right]^2} \left[\dot{x}_n(t) - \dot{x}_{n+1}(t)\right] \tag{49}$$

which can be integrated to give us

$$u = u_f \; exp \left(-\frac{k}{k_{\text{jam}}} \right) \tag{50}$$

It should be emphasized that the pursuit of more accurate car-following models is always rooted in the formal hypothesis of (44), which assumes that the follower's acceleration is a function of speeds, relative speeds, and spacings. This component of the model goes beyond physical considerations, since it tries to replicate human behavior, thereby increasing the complexity of the vehicle-driver system.

To conclude this non-exhaustive overview of car-following models based on the stimulus-response modeling hypothesis, two notable models that follow a similar trajectory are the Gazis-Herman general model and the Ahmed model. The Gazis-Herman model, proposed by Gazis et al. (1961) is given by:

$$\ddot{x}_{n+1}(t+\tau) = \lambda \frac{\left[\dot{x}_{n+1}(t)\right]^m}{\left[x_n(t) - x_{n+1}(t)\right]^l} \left[\dot{x}_n(t) - \dot{x}_{n+1}(t)\right] \tag{51}$$

where l and m are parameters without physical meaning (in order to better fit the observations).

A generalized version of this model has been proposed by Ahmed (1999) and assumes an acceleration rate given by:

$$\ddot{x}_{n+1}(t+\tau) = \alpha^{\mp} \frac{\dot{x}_{n+1}^{\beta^{\mp}}}{g_{n+1}^{\gamma^{\mp}}} \left[\dot{x}_n(t) - \dot{x}_{n+1}(t)\right] \tag{52}$$

These models assume different driver behaviors in following vehicles, depending on whether they are in the acceleration or braking phase. The model parameters $\alpha^+, \beta^+, \gamma^+$ are used for acceleration when $\dot{x}_n(t) \geq \dot{x}_{n+1}(t)$, and $\alpha^-, \beta^-, \gamma^-$ are used for deceleration when $\dot{x}_n(t) \leq \dot{x}_{n+1}(t)$. Here, l_n is the vehicle's length, and $g_{n+1} = x_n(t) - x_{n+1}(t) - l_n$ represents the gap distance from the leading vehicle (sometimes called the "effective distance").

In parallel to these developments other researches have sought a unified functional framework for car-following models. Newell (1961) explicitly expresses the dynamics in terms of a velocity-headway function:

$$\dot{x}_{n+1}(t) = V\left\{1 - exp\left[-\frac{\lambda}{V}\left(x_n(t) - x_{n+1}(t)\right) - d\right]\right\} \tag{53}$$

where λ and d are constants (calibration parameters) and V is the top velocity. Bando et al. (1995), Bando et. al (1998), and Treiber and Kesting (2013) propose a variant known as the "Optimal Velocity Model" (OVM), based on the dynamic equation:

$$\ddot{x}_{n+1}(t+\tau) = \alpha\left[V\left(s_{n+1}\right) - \dot{x}_{n+1}(t)\right] \tag{54}$$

Here, α is an acceleration constant, τ is a time-lag (which could represent the reaction time), and $V\left(s_{n+1}\right)$ is a velocity-headway function with s_{n+1} as the headway. One of the most widely used models within this family is the Intelligent Driver Model (IDM) by Kesting, Treiber and Helbing (2010):

$$\ddot{x}_{n+1}(t+\tau) = a\left\{1 - \left(\frac{\dot{x}_{n+1}(t)}{v_0}\right)^{\delta} - \left[\frac{s^*\left(\dot{x}_{n+1}(t), \Delta v_{n+1}(t)\right)}{s_{n+1}(t)}\right]^2\right\} \tag{55}$$

and

$$s^*\left(\dot{x}_{n+1}(t), \Delta v_{n+1}(t)\right) = s_0 + \dot{x}_{n+1}(t)T + \frac{\dot{x}_{n+1}(t)\Delta v_{n+1}(t)}{2\sqrt{ab}} \tag{56}$$

Here, T represents an anticipation time that also takes into account the velocity difference $\Delta v_{n+1}(t) = v_{n+1}(t) - v_n(t) = \dot{x}_{n+1}(t) - \dot{x}_n(t)$, that is, the approaching rate to the leading vehicle. The IDM combines a free flow acceleration function of the speed $a_f(v) = a\left[1 - \frac{v}{v_0}\right]^{\delta}$ with a braking strategy to decelerate $a_b = -a\left(\frac{s^*}{s}\right)^2$. This strategy becomes relevant when the gap between the follower and the leader is not significantly larger than the "effective gap" $s^*(v, \Delta v)$. The desired speed, v_0, is a behavioral parameter that differentiates drivers, and maximum acceleration is denoted by the parameter a, which also allows differentiating vehicles, meaning that the vehicles are not clones of each other. Finally, parameter δ characterizes how the acceleration changes with speed. Parameters a and b can be measured and calibrated (Kesting et al., 2010).

Ward and Wilson (2011) and Wilson (2011) formulate a common functional framework for car-following models, where a follower's reaction in terms of acceleration or deceleration depends on speeds, spacings, and relative speeds:

$$a_{n+1}(t) = \ddot{x}_{n+1}(t) = \mathcal{F}\left[s_{n+1}(t), \Delta v_{n+1}(t), v_{n+1}(t)\right] \tag{57}$$

These models have "uniform flow" steady solutions (equilibria) if, for each $s^* > l$, there is a $v^* = V(s^*) > 0$ such that $\mathcal{F}(s^*, 0, v^*) = 0$. Here, $V(s^*)$ represents the equilibrium speed-spacing relationship that leads to a fundamental diagram, thus ensuring that the car-following model is consistent with traffic flow theory. The general functional approach $\mathcal{F}$ serves a special interest by also laying the foundation for analyzing the stability of car-following models in terms of the partial derivatives, as outlined by Treiber and Kesting (2013). These derivatives must satisfy the conditions:

$$\mathcal{F}_s > 0,\ \mathcal{F}_{\Delta v} > 0,\ \mathcal{F}_v < 0 \tag{58}$$

To conclude this summary overview of these car-following models as alternatives to modeling the behavior of traffic flows, we will deal strictly with the main modeling aspects and discuss the family known as "collision avoidance" models (Gerlough and Huber, 1975; Barceló, 2010). These models assume that a follower driver will attempt to maintain a safety distance $s_n(t)$ from the lead vehicle, such that in the event of an emergency stop by the leader, the follower will come to a stop without colliding with the lead vehicle. The safe deceleration-to-stop diagram in Figure 13 (Gerlough and Huber, 1975; and Mahut, 1999) illustrate how this concept works.

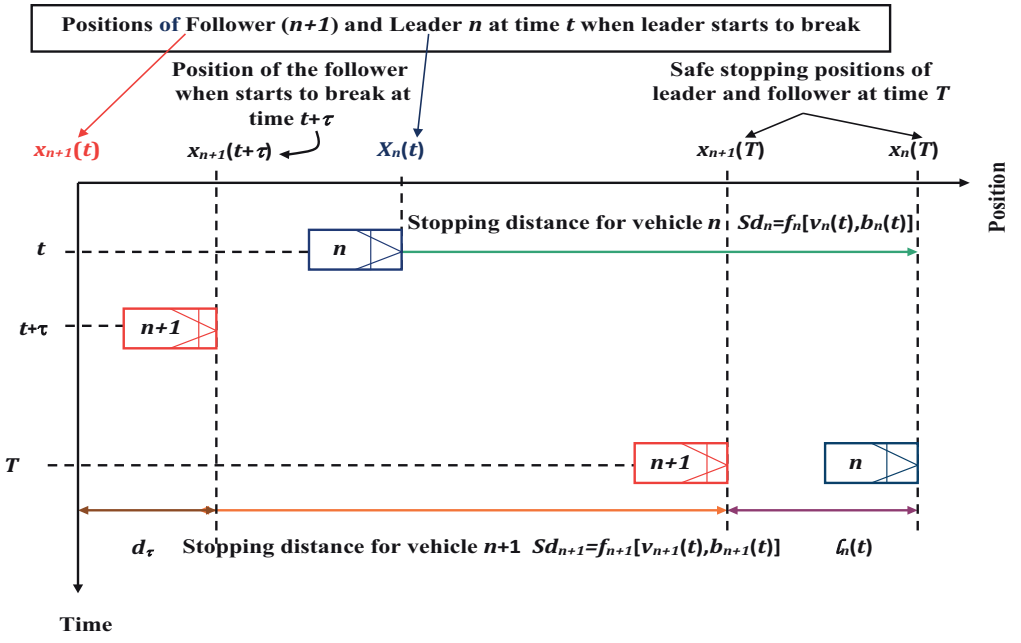

Figure 13. *Safe to stop diagram.*

This time-space diagram, as discussed in Barceló (2010), shows the positions of the leader n at time t, beginning with the initiation of braking until coming to a complete stop at time T. The follower $n+1$ perceives the leader's braking with a delay τ, representing the reaction time. During this delay, the follower travels a distance d_τ, which covers the

position from the initiation of braking to coming to a safe stop. If Sd_n is the stopping distance for vehicle n, that is, the distance required to stop when traveling at speed $v_n(t) = \dot{x}_n(t)$ at time t, and the driver brakes with a deceleration function $b_n(t)$, then $Sd_n$is given by $Sd_n = f_n[v_n(t), b_n(t)]$. This function depends on the current speed $v_n(t)$ and the applied deceleration $b_n(t)$. Similarly, the stopping distance for the follower vehicle is given by $Sd_{n+1} = f_{n+1}[v_{n+1}(t), b_{n+1}(t)]$. Then, the desired spacing $s_n(t) = x_n(t) - x_{n+1}(t)$ at time t for a safe deceleration-to-stop is given by:

$$s_n(t) = x_n(t) - x_{n+1}(t) = \dot{x}_{\mathbf{n+1}}(t) \cdot \tau + Sd_{n+1}[v_{n+1}(t), b_{n+1}(t)] + \ell_n(t) - Sd_n(t)[v_n(t), b_n(t)] \tag{59}$$

where $d_\tau = \dot{x}_{\mathbf{n+1}}(t) \cdot \tau$ is the distance traveled by the follower during the reaction time τ, and $\ell_n(t)$ is the minimum safety distance (i.e., the distance between bumpers at rest). Assuming steady-state conditions with equal deceleration functions $b_n(t) = b_{n+1}(t)$, equal speeds, and, therefore, $Sd_{n+1} = Sd_n$, we have:

$$x_n(t) - x_{n+1}(t) = \dot{x}_{\mathbf{n+1}}(t+\tau) \cdot \tau + \ell_n(t)$$

Differentiating with respect to t, we obtain:

$$\dot{x}_n(t) - \dot{x}_{n+1}(t) = \tau \ddot{x}_{n+1}(t+\tau) \;\rightarrow\; \ddot{x}_{n+1}(t+\tau) = \frac{1}{\tau}[\dot{x}_n(t) - \dot{x}_{n+1}(t)]$$

which is the elementary form of the response to a stimulus model (45). Rewriting (59) as

$$x_n(t) + Sd_n(t)[v_n(t), b_n(t)] - \ell_n(t) \geq x_{n+1}(t) \dot{x}_{\mathbf{n+1}}(t) \cdot \tau + Sd_{n+1}[v_{n+1}(t), b_{n+1}(t)] \tag{60}$$

we can interpret this equation as a safety constraint that becomes active when it is satisfied as an equality, thus activating the braking action from the follower. Assuming steady-state conditions with $\ell_n(t) = \ell_n$ and constant deceleration functions $b_n(t)$ and $b_{n+1}(t)$ for a period, the respective distances to stop can be expressed as:

$$Sd_n(t) = -\frac{\dot{x}_n^2(t)}{2b_n} \;\; and \;\; Sd_{n+1}(t) = -\frac{\dot{x}_{n+1}^2(t+\tau)}{2b_{n+1}} \tag{61}$$

From equation (60), Mahut (1999) and Barceló (2010), we replace the Gipps speed of the follower during the deceleration phase, denoted as $v_{n+1}^b(t+\tau)$ (Gipps, 1981) and can thus be derived as:

$$v_{n+1}^b(t+\tau) = b_{n+1}\tau \sqrt{b_{n+1}^2 \tau - b_{n+1}\left[2\{x_n(t) - \ell_n - x_{n+1}(t)\} - v_n(t)\tau - \frac{(v_n(t))^2}{b_n}\right]} \tag{62}$$

This includes a change proposed by Gipps (1981), which is based on empirical and consistency analyses. It replaces the leader's braking deceleration $b_n$with an estimated

value $\widehat{b}_n$ because the follower does not have precise knowledge of the leader's deceleration. Furthermore, a safety margin is included to allow the follower a possible delay θ (e.g., $\theta = \frac{\tau}{2}$) when traveling at $v_{n+1}(t+\tau)$ before reacting to the braking of the vehicle ahead, thus satisfying:

$$-\frac{[v_{n+1}(t+\tau)]^2}{2b_n} + v_{n+1}(t+\tau)\left(\frac{\tau}{2}+\theta\right) - [x_n(t) - \ell_n - x_{n+1}(t)] + \frac{v_{n+1}(t)\tau}{2} + \frac{(v_n(t))^2}{2\widehat{b}_n} \leq 0$$

The so-called microscopic traffic models described in Barceló (2010) aim to replicate the propagation of traffic flows in a road network. The process, known as network loading, models the dynamics of car-following, lane-changing, and gap acceptance as vehicles travel from origins to destinations following route choice algorithms that mimic drivers' decisions. Microscopic traffic models have an advantage over continuum flow models because they handle traffic flow interruptions naturally, since car-following explicitly accounts for the possibility of the leader stopping. Consequently, these models can explicitly include traffic lights at signalized intersections to accurately represent the detailed phasing and timings. Figure 14 provides a graphical summary of the model-building process and its components.

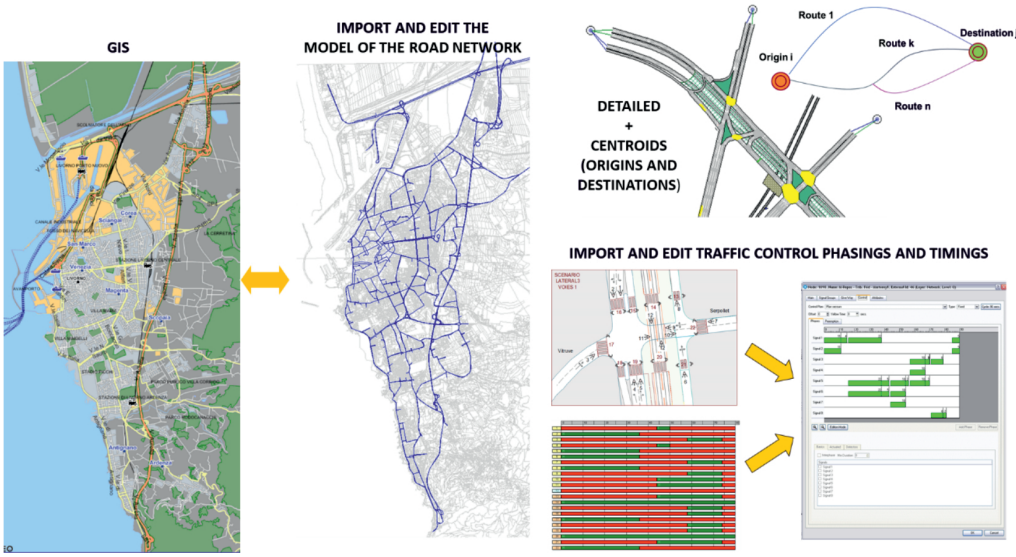

Figure 14. *Scheme of the microscopic model building process and the model components.*

2.5. Mesoscopic traffic models

DTA models have always been appealing due to two key characteristics. Firstly, they can handle large road networks, similar to the static assignment models used in transport planning. Secondly, they can account for time variations in transport demand and their impacts on the road network. However, their initial analytical approaches to solving

the DTA formulation in (21), (22), and (23) proved to be computationally challenging. This generated interest in exploring alternative traffic simulation-based approaches that can provide approximate heuristic solutions. Florian, Mahut and Tremblay (2001) and Florian, Mahut and Tremblay (2002) proposed a conceptual framework that integrates both analytical and simulation-based approaches, as illustrated in Figure 15 (Barceló, Ros-Roca and Montero, 2022). The framework consists of two main interdependent components:

- A method for determining path-dependent flow rates on the network paths, which can be approached using various algorithms, ranging from the exact projection methods mentioned earlier to approximations like the Method of Successive Averages.

- A Dynamic Network Loading (DNL) method, which determines how these path flows translate into time-dependent arc volumes, arc travel times, and path travel times.

In the most successful practical implementations, the DNL method is usually based on a mesoscopic simulation model (Barceló, 2010) that emulates the flow propagation through the network under the current conditions. The resulting assignment depends on how the convergence criterion and iterative process are implemented. It can be a DTA or a dynamic user equilibrium (DUE) (Chiu et al., 2011). A mesoscopic traffic simulation model of traffic flow dynamics is a simplified representation that captures key aspects of microscopic simulation while being less data demanding and computationally more efficient than microscopic models. Mesoscopic approaches combine microscopic and macroscopic aspects of traffic flow dynamics, providing an alternative approach depending on the modeling objectives and hypotheses. In this paper, we will focus on approaches where flow dynamics are determined by the simplified dynamics of individual vehicles.

One such approach is the *cell transmission model* proposed by Daganzo (1994) and (1995a), which solves an ad hoc version of the first order traffic flow model using a simplified flow-density relationship known as the triangular (or trapezoidal) fundamental diagram (Daganzo, 1995b).

This basic model and its many variants, although widely used, exhibit limitations, namely in the case of urban networks, since they only account for flow dynamics in links and do not explicitly deal with intersections, most notably the signalized intersections commonly found in urban networks. To overcome these drawbacks, various extensions have been proposed to incorporate intersection modeling. One notable example is the *general link transmission model* (GLTM) developed by Gentile (2010) and Gentile (2015).

Another modeling alternative involves splitting the link into two parts, as shown in Figure 16. The first is the running part, where vehicles are not yet delayed by the queue spillback at the downstream node. The second is the queueing part, where the capacity is

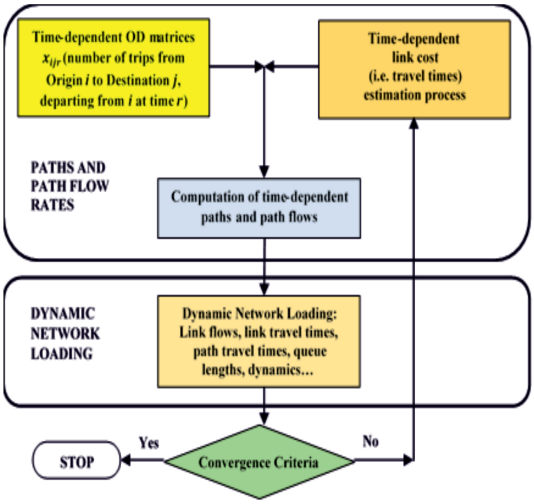

Figure 15. *Conceptual algorithmic framework for DTA.*

limited by stop signs, give-way signs, or traffic lights. Nodes are the interactions between traffic flows at intersections, and they can be modeled using either node transfer modules or a queue server approach that explicitly considers traffic lights and the delays they cause (Mahmassani et al., 1994). The flow dynamics in the running part are simplified in terms of macroscopic speed-density relationships by using variants of (36), such as:

$$u_i^t = (u_f - u_0)\left(1 - \frac{k_i^t}{k_{\text{jam}}}\right)^{\alpha} + u_0 \tag{63}$$

This equation, proposed by Jayakrisham, Mahmassani and Yu (1994), relates the mean speed u_i^t and density k_i^t in section i at time step t. The parameters u_f and u_0 are the mean free speed and the minimum speed, k_{jam} is the jam density, and α is a parameter that captures speed sensitivity to density. Alternatively, Ben-Akiva et al., (2001), and (2010), propose the following speed-density relationship:

$$u = \begin{cases} u_f & \text{if } k < k_{\text{Min}} \\ u_f\left[1 - \left(\frac{k - k_{\text{Min}}}{k_{\text{jam}}}\right)^{\alpha}\right]^{\beta} & \text{otherwise} \end{cases} \tag{64}$$

including a lower bound limiting density, k_{Min}, and a second parameter β to capture speed sensitivity to concentration. Vehicle dynamics in the queueing part are then governed by the queue discharging process. The boundary between the running part and the queueing part is dynamic and varies according to the queue spillback and queue discharge processes.

A completely different approach is taken by Mahut and Florian (2010), who propose a simulation model that moves vehicles individually using a simplified car-following model. In this model, the position $x_{n+1}(t)$ of the follower vehicle $(n+1)$ at time t relative

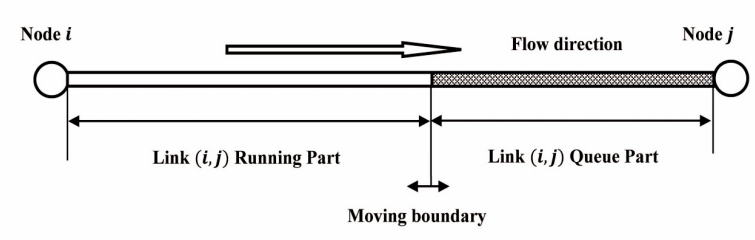

Figure 16. *Link model.*

to the position $x_n(t-T)$ of the leader vehicle (n) is estimated according to:

$$x_{n+1}(t) = \text{Min}\left[x_{n+1}(t-\varepsilon) + \varepsilon u_f, x_n(t-T) - l_{eff}\right] \tag{65}$$

where T is the reaction time, u_f the free-flow speed, l_{eff}, is the effective vehicle length, and ε is an arbitrarily short time interval. The first term inside the minimizing operator represents the farthest position downstream that the vehicle can reach at time t based on the follower's position at time $(t-\varepsilon)$ under the constraints of the maximum speed u_f. The second term inside this operator represents the farthest downstream position that can be attained based on the trajectory of the next vehicle downstream in the same lane, according to a simple collision-avoidance rule proposed by Mahut (1999) and Mahut (2001). This simplified model depends only on the free-flow speed and does not account for accelerations. It can be considered a lower-order model, since it determines the position of each vehicle in time, rather than their speed or acceleration.

The solution to the car-following relationship (65) for time can be expressed as:

$$t_{n+1}(x) = \text{Max}\left[t_{n+1}(x-\delta) + \frac{\delta}{u_f}, t_n(x+l_{eff}) + T\right] \tag{66}$$

This relationship allows for an event-based simulatiaon-4mm approach, as it enables calculating the link entrance and exit times for each vehicle by means of the following expression in Equation (67):

$$t_{n+1}(L_1) = \text{Max}\left[t_{n+1}(0) + \frac{L_1}{u_f^1},\ t_n(L_1) + T + \frac{l_{eff}}{\text{Min}\left[u_f^1, u_f^2\right]}, t_{n+\frac{L_2}{l_{eff}}}(L_2) + \frac{L_2}{l_{eff}}T\right] \tag{67}$$

where L_1 and L_2 are the lengths of two sequential links with speeds u_f^1 and u_f^2, respectively. The vehicle attributes l_{eff} and T are assumed to be identical throughout the entire traffic stream, and each vehicle adopts the link-specific free-flow speed when traversing a given link. The link lengths are assumed to be integer multiples of the vehicle length, l_{eff}. As shown by Mahut (2000), this model yields the triangular fundamental flow-density diagram proposed by Daganzo (1994). The main events that change the state of the model include vehicle arrivals at links, link departures, and transfers between links based on turning movements at intersections.

2.6. Models, valid models, and the data requirements for validating transport models

There is a common argument that emphasizes the abundance of data and suggests that models are becoming less necessary. Proponents of this view often strengthen their argument by quoting George Box's statement, "All models are wrong." Box, who is considered to one of the founding fathers of modern statistics and an expert in modeling recognizes, seemingly implies that models are useless. Although Box indeed made that assertion in 1976, the full context is often overlooked by those who quote him, as he later added an important caveat in the book by Box and Draper (1987), where he added "but some are useful," which, to the best of my knowledge, is frequently omitted.

Models as formal representations of systems are only approximations. As such, one should never forget that a representation of a system is not the actual system. Furthermore, when taking the Minsky triad perspective of building a model of a system, the modeler must ensure they are asking the right question and that the modeling hypotheses align with the objectives. As demonstrated in the previous sections discussing various modeling alternatives for traffic and transport systems, the goal is to highlight the diverse options available. Hence, considering Box's statement that some models are useful, the key question becomes: What makes a model useful?

The answer to that question is proper model validation and calibration, which are defined by Barceló (2010) and MULTITUDE (2014) as the following.

- **Calibration** is the process of determining model parameter values based on field data in a specific context. Parameters for the transport model in one city will differ from those for another city, and the nature of the parameters depends on the type of model and the objectives of the decision-making process supported by the model

- **Validation** aims to provide a quantitative answer to the question of whether model predictions faithfully represent reality. According to Rouphail and Sacks (2003), validation is determined formally by the probability that the difference between the "reality" and the "model prediction" falls within a tolerable difference threshold, denoted as d. This threshold measures the model's proximity to reality or, in other words, the error incurred when substituting the reality with the model. The level of assurance, denoted as a, measures the degree of certainty when making this substitution. The validation process satisfies the following criterion:

$$P\{|\text{reality} - \text{“model prediction”}| < d\} > a \tag{68}$$

It is the responsibility of the modeler conducting the study to define the criteria for model validation and acceptance. These criteria determine the values of parameters a and d, which assess the suitability and acceptability of the model. Figure 17 summarizes the methodological processes for calibrating and validating models, along with the data requirements for these processes. Looking at equation (68), in essence, the validation and calibration processes involve a statistical comparison between the observed "reality"

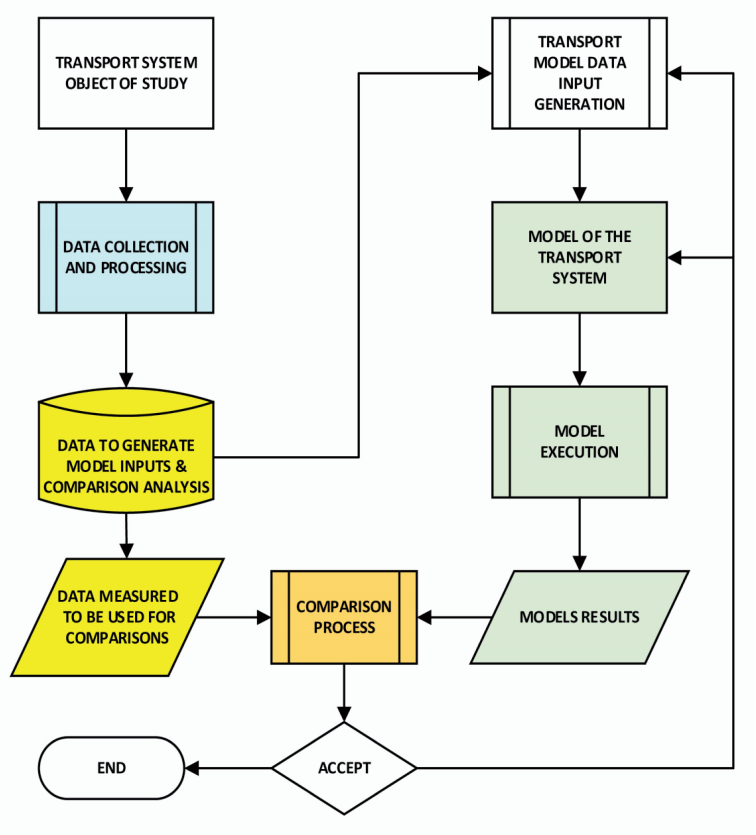

Figure 17. *Methodological scheme of model calibration and validation processes.*

and "predicted" values of relevant variables that define the state of the system (e.g., speeds, flows, travel times). Usually, after identifying the variables to be used in the process, the corresponding data must be collected and processed using appropriate data analysis techniques. This includes cleaning the data, removing outliers and erroneous measurements, mitigating biases, and addressing missing values. Samples of observed data are usually split into independent subsets for various uses, such as calibrating model parameters, conducting statistical comparisons between observed and predicted data to validate the model, and, in some cases, generating inputs for the model.

All the models mentioned so far require the determination of parameter values and the calibration of data inputs, which are highly dependent on the specific system under study. Below are some examples.

- Static traffic assignment models:
  - BPR volume-delay functions for each link a (or class of links): t_0^a free flow travel time, κ_a link capacity, function parameters α_a, β_a
  - Conic functions, the additional delay factor J_a

  - $\Delta = [\delta_{ap}]$ the link-path incidence matrix that depends on the network topology
  - $X = [X_{rs}]$ number of trips from origin r to destination s for each OD pair (r,s)

- Dynamic assignment models

  - Time-dependent path cost functions $tt_{rsp}(t)$
  - Time-dependent (dynamic) OD matrices: $X(t) = [X_{rs}(t)]$

- Traffic flow theory models. The models described earlier involve various sets of parameters, depending on the order of the model:

- First-order models: the main parameters to calibrate determine the fundamental diagram, such as free flow speed u_f, the jam density k_{jam}, and the parameters *?* and *?* in equation (36).

- The second-order models in (38) will additionally require estimations of the equilibrium speed u_e and the viscosity v.

- Microscopic traffic models. These depend on a vast number of parameters and, as already mentioned, many aspects of the model-building process is automated in terms of importing road maps from GIS and setting the controls using original files, to name but two examples. For the purposes of this paper, let us focus here on car-following models and the parameters of the simulation engine:

- The Herman-Gazis car-following model (52): Parameters to fine-tune include values for the gap distance g_n between the leader and the follower as a function of the leader's length ℓ_n, as well as the acceleration and deceleration parameters $\alpha^{\mp}$, $\beta^{\mp}$, $\gamma^{\mp}$.

- The Gipps car-following models (62): Parameters to consider include deceleration rates b_n, reaction time τ, adjustment factor θ, and others.

It is worth noting that other car-following models, such as IDM (Kesting et al., 2010), the Wiedemann psycho-physical car-following model (Wiedemann, 1974), Fritzsche's car-following model (Fritzsche, 1994), or Krauss's car-following models (Krauss, Wagner and Gawron, 1997) are among the most widely used in the current traffic simulation software platforms.

Additionally, microscopic traffic simulation models, which involve vehicles travelling across the network from origins to destinations along paths, share certain requirements with DTA models. These include the calibration of time-dependent link travel times, time-dependent OD matrices $X(t) = [X_{rs}(t)]$, and path choice models that are typically based on discrete choice theory. Their utility functions may depend on factors like the value of time, which needs to be calibrated. Although controversial, path choice

models are essential in capturing the behavioral aspects and topology of transportation modeling, especially in urban networks where the phenomenon of sharing links is common. To illustrate the nature of the problem, let us consider the following notation: $K_{rs}(t)$ is the set of paths from origin r to destination s at time t, $p(r,s,t)$ denotes the path $p \in K_{rs}(t)$, and $\Gamma_{p(r,s,t)} = \{e_1, \ldots, e_m\}$ is the set of links of path $p(r,s,t)$. If l_a is the length of link a, and $L_{p(r,s,t)}$ is the total length of path p, the commonality factor (Cascetta et al., 1996; Ben Akiva and Bierlaire, 1999) is a measure of the fraction of path p that is shared with all other paths h connecting origin r with destination s at time interval t. It is given by:

$$CF_{p(i.j.t)} = \frac{1}{\mu_{CF}} \sum_{a \in \Gamma_{p(,i,j,t)}} \left(\frac{l_a}{L_{p(i,j,t)}} \log \left(\sum_{h \in K_{rs(t)}} (\delta_{aht} + 1) \right) \right) \tag{69}$$

where δ_{aht} indicates whether link a also belongs to path $h \in K_{rs(t)}$ or not. The path choice proportion $P_{p(i,j,t)}$for each path on the set $K_{rs}(t)$ is calculated as a discrete choice model that uses the commonality factor within the OD pair and time. $CF_{p(r,s,t)}$ is a penalization factor added to current travel times (Bovy, Bekhor and Prato, 2008; Janmyr and Wadell, 2018):

$$P_{p(r,s,t)} = \frac{exp[\mu_{P_p}(-\widehat{tt}_{p(r,s,t)} - CF_{p(r,s,t)})]}{\sum_{h \in K_{rs(t)}} exp[\mu_{P_p}(-\widehat{tt}_{h(r,s,t)} - CF_{h(r,s,t)})]} \tag{70}$$

where $\widehat{tt}_{p(r,s,t)}$ is either the average travel time on the path $p \in K_{rs}(t)$ or the estimates of the average travel time value, and μ_{CF} and μ_{P_p} are parameters that must be calibrated.

- Mesoscopic traffic models depend on the set of values identified in equations (65)–(67), which must obviously be calibrated. Their relationships with the triangular fundamental diagram proposed by Daganzo (1994) must be explicitly considered. However, looking at the methodological computational scheme in Figure 15, we can see that they require inputs such as time-dependent OD matrices $X(t) = [X_{rs}(t)]$ and path choice functions similar to equations (68) and (69), which they have in common with the microscopic simulation models.

3. Data collection and analysis

The introductory remarks in this section emphasize the interdependence between data and models. It is essential to have data for the model-building process, and the overall model-building and utilization processes are illustrated in Figure 1. In Section 2.6, the conditions that render models useful are established through the model calibration and validation processes, which also need data, as depicted in the methodological scheme in Figure 17. Figure 18 summarizes a methodological approach that combines the key concepts outlined in Figures 1 and 17, incorporating elements from the OECD/ITF report (2015).

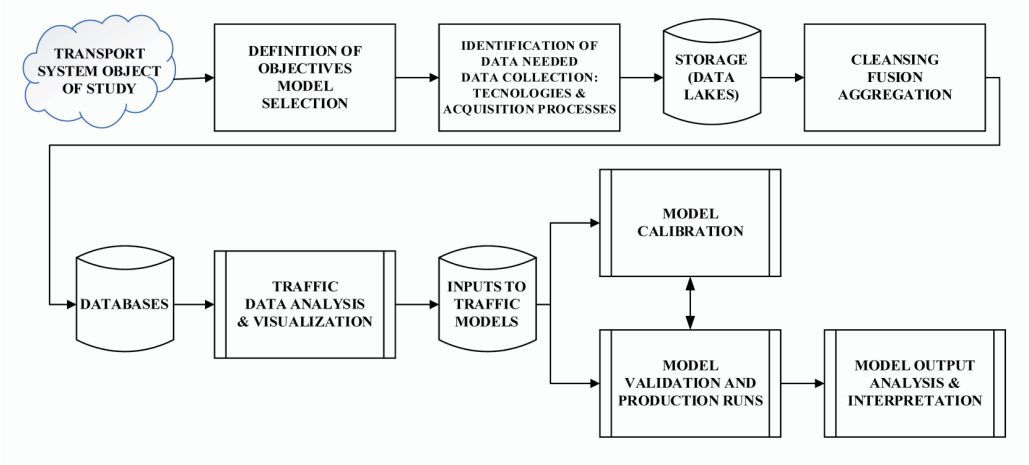

Figure 18. *Data collection and processing: A methodological approach for model construction, calibration, and validation.*

The steps in the process, which appear in the boxes of the methodological diagram are the following:

A. The first step, outlined in Section 1 of this paper, establishes a fundamental methodological principle for using models in analyzing transport systems. It emphasizes that different modeling approaches can be employed based on the objectives and associated modeling hypotheses. Therefore, the first step in using models to analyze transport systems is to identify the study objectives and determine the most suitable modeling approach to achieve them.

B. The second step, described in Section 2 of this paper, provides an overview of the most common modeling approaches for transport systems, depending on the objectives. These range from strategic approaches that employ either static or dynamic traffic assignment to the macro, meso, or micro approaches suitable for operational analysis or other purposes. Section 2.6 identifies the main parameters that must be calibrated for proper model building and use, along with the data needed for that process. Once the necessary data have been identified, the analysis must also ascertain the available technologies (Guerrero-Ibáñez, Zeadally and Contreras-Castillo, 2018) for data collection and determine the appropriate procedures.

B.1 Point detection with discrete time resolution: This includes inductive loop detectors, radars, etc., placed at specific positions, as depicted in Figure 19, which also indicates whether they are single or double loop detectors. They provide aggregated measures, with a Δt time resolution of:

- Average traffic flows in vehicles/hour

- Average occupancies: percentage of time a vehicle is over the detector with respect to the aggregation time Δt
- Average spot speeds (speeds measured at the detection point) in km/hour
- Traffic mix: percentages of light and heavy vehicles

B.2 Point detection with continuous time resolutions: This also occurs at specific positions, either on the road (e.g., inductive loops like magnetometers, as illustrated in Figure 19) or as roadside units (e.g., Bluetooth/Wi-Fi antennas, electronic TAG readers, CCTV image processing, etc.).

- Magnetometers measure the time in / time out of a vehicle passing over the detector and provide count flows, spot speeds, occupancies, and traffic mix, which can also be aggregated accordingly.
- Bluetooth/Wi-Fi and TAG readers identify the corresponding device onboard the vehicle and reidentify it downstream. They count only the flows of equipped vehicles (a non-representative sample of the whole population) and the time differences between two successive detection devices. Considering that the positions are well known and time differences are highly accurate, they provide a good sample of travel times or speeds between specific pairs of locations.
- CCTV image processing devices located at specific positions identify a vehicle by reading its license plate (license plate recognition) and reidentify it downstream. They have the advantage of detecting all vehicles, allowing measurement of point traffic flows and travel times between camera locations. If properly located, the cameras can also provide an estimate of OD matrices, with the origin being the point where the vehicle is first detected and the destination where it is last detected.

B.3 Continuous time-space detection, enabled by mobile devices that can be tracked along their trajectories, provides:

- In the case of GPS devices: waypoints with the detection time-tag; the vehicle's location (automatic vehicle location, AVL), consisting of the x, y and z coordinates (as shown in Figure 19); and, if the mobile device allows, the point speed at the detection time and heading.
- Mobile phones provide data from the call detail records, which can be processed to extract movements between origins and destinations. In some cases, inferences can be made about the routes used.
- Connected vehicles provide similar information about origins-destinations, travel times, speeds, locations, etc., either directly or via roadside units. In some cases, they can also provide additional similar data about the surrounding cars.

B.4 Public transport: Contactless cancellations provide a rich amount of data about passenger usage and transfers in public transport. Additionally, other ICT applications can provide passenger counting data aboard vehicles, etc. Bus monitoring systems provide detailed tracking information on schedule adherence, bus speeds, arrival times at stops, etc.

B.5 Shared vehicles: Data recorded by shared services vary depending on the company, type of vehicle (car, bike, etc.), and the equipment installed on the vehicle. Currently, there is no standard common type of data recorded. What is recorded may range from time and location of the service's origin and end, and in some cases a track of the route used.

Note: One subject of intense research has been optimizing the placement of point detectors, inductive loops, magnetometers, CCTV cameras, Bluetooth antennas, and other devices to provide measurements of partial path travel times, OD estimation, and others. This interesting optimization problem concerns the coverage problem in networks and the observability of the traffic system when measurements are used to estimate the system's state. Although the analysis of these models is beyond the scope of this paper, interested readers can find comprehensive overviews in Barceló et al. (2012) and Castillo et al. (2008).

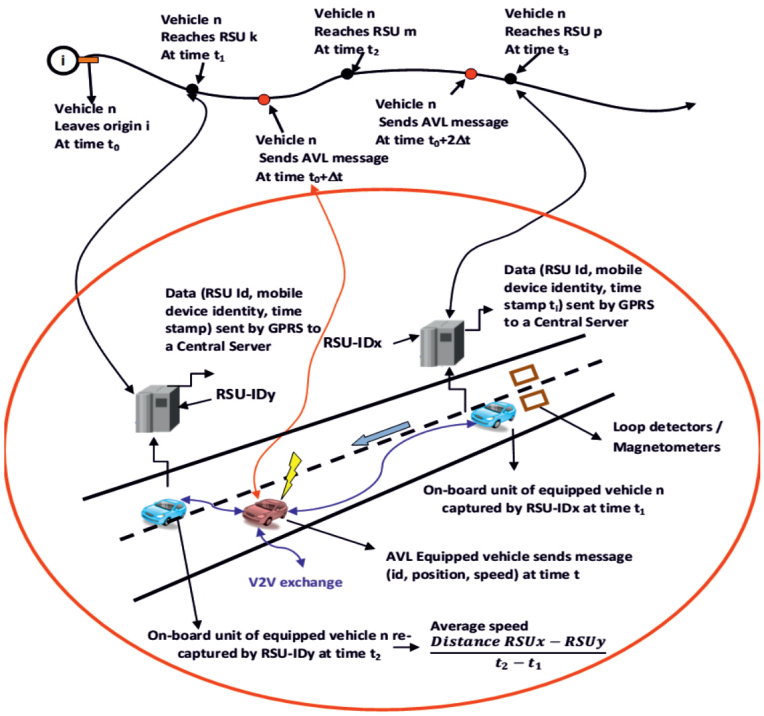

Figure 19. *Examples of data collection technologies.*

B.6 Other data sources

- Other ICT sources, such as social networks (e.g., WIZE, Moovit) or Google, can also provide data. However, these data often require specific treatment before being used in conjunction with other sources.

- Non-ICT networks: The summary overview of transportation modeling approaches highlights the key role played by traffic analysis zones (TAZ) and modal splitting in the four-step modeling approach. These models typically require socioeconomic and other data, such as information of transportation mode usage. No such data are currently available from TIC applications, although research efforts are actively addressing this gap. Furthermore, when dealing with transport mobility patterns represented by OD matrices, the usual ICT data sources only provide partial samples (e.g., from a subpopulation of vehicles equipped with a specific technology) that cannot be extrapolated to the whole population without complementary data. These complementary data sources are usually census tracts or carefully designed household surveys.

C. Third step: conventional data collection technologies such as those based on magnetic loops have historically provided limited and frequently scarce point observations at detection station locations. However, the emergence of new information and communication technologies (ICT) has dramatically changed the situation by granting access to large volumes of data, primarily spatial data, which necessitated proper storage using big data techniques (OECD/ITF, 2015). This vast and heterogeneous raw data is initially stored in unstructured data lakes, with an emphasis on acquiring it quickly, particularly for real-time operations.

D. Fourth step: Cleansing, fusion, and aggregation. Regardless of the quality of detection technologies, they are nevertheless prone to errors induced by temporal detection malfunctions or external factors affecting the detection quality (e.g., limited accuracy of GPS signals in certain urban areas). Therefore, the collected data must be properly cleansed before being used in transport models. This involves a series of data processes for identifying and filtering outliers to mitigate the risk of inducing undesirable biases, as well as completing missing data caused by outlier removal or lack of detection during a certain period. The objective of this step is to get clean, consistent, and complete data series.

   However, these clean and consistent data cannot be directly utilized since they originate from diverse data sources. For example, speeds measured by Bluetooth, CCTV cameras and inductive loop detectors may need to be fused. Similarly, generating modal split distributions may require combining data from mobile phones and household surveys. Data fusion techniques are employed to homogenize and harmonize these heterogeneous datasets to generate unique and consistent inputs.

Different models, depending on their specific use, may require different levels of data aggregation. For example, when dealing with OD matrices, time aggregation differs between static assignment models, dynamic assignment models, and dynamic traffic models. Therefore, depending on the model requirements, the data must be appropriately aggregated.

E. Fifth step: These clean, consistent, and structured datasets are then stored in databases that are specifically designed for data retrieval, tailored to the needs of different transport models.

F. Sixth step: Traffic data analysis and visualizations.

- This step involves generating the inputs required by the transport models, which will be elaborated upon in the subsequent sections.
- Additionally, advancements have been made in graphical techniques, which allow for data visualization and heat maps depicting degrees of congestion on network links or paths. These advances also highlight the attraction and generation capacities of the TAZs, such as in depicting travel patterns from origin to destination. Such descriptions of the system are useful for understanding the state of the system and assisting in the decision-making process.

G. Seventh step: Input to traffic models. The data for the transport model used in the study must formatted appropriately as input for the subsequent steps: calibration, validation and, once the model is validated, the production runs corresponding to the various scenarios to be analyzed and compared for decision-making purposes.

3.1. Notes on data cleaning

To illustrate some of the techniques used for data cleaning, including outlier removal, replacement of missing values, and correction of erroneous values, we will discuss two specific cases: the application of the Kalman filter to handle series of travel time measurements between two consecutive Bluetooth antennas; and using a map matching process to determine the correct position of a GPS waypoint within a network link.

3.1.1. Using a Kalman filter to clean time series of bluetooth travel time measurements

The Kalman filter, introduced by Kalman (1960) and further developed by Dailey, Harn and Lin (1996), is a state space model used to estimate the dynamics of a system. In this model, the state of the system at time instant k is defined by a set of unobservable state variables, represented by the vector $x_k \in R^p$ (where p is the number of state variables). The evolution of the system state transitions over time is governed by the linear stochastic equation in differences:

$$x_k = \Phi \cdot x_{k-1} + w_k \tag{71}$$

where Φ is the transition matrix and w_k represents the process noise, which is assumed to be white Gaussian noise with zero mean and covariance matrix Q. The system is observed at time k with measurements denoted as $y_k \in R^q$ (where q is the number of observations). The relationship between the measurements and the state variables is given by the linear measurement equation:

$$y_k = A \cdot x_k + v_k \tag{72}$$

where A is the measurement matrix with measurement noise v_k, which we also assume to be white Gaussian noise with zero mean and covariance matrix R. The process and measurement noises are assumed to be independent, with covariance matrices Q and R that can change at each step. The discrete Kalman filter cycles recursively between a temporal update and an estimation step. The temporal update projects the immediate future of the current state, and the covariance estimation provides an a priori estimate from steps k-1 to k, all by means of the following:

$$\begin{aligned} \widehat{x_k^-} &= \Phi \cdot \widehat{x}_{k-1} \\ P_k^- &= \Phi \cdot P_{k-1} \cdot \Phi^{\mathsf{T}} + Q \end{aligned} \tag{73}$$

The measurement update adjusts the projection of the estimate by incorporating the measurements available at that moment. It begins by calculating the Kalman gain, G_k, which is used to generate a posteriori estimates by incorporating the measurements $y_k$at that instant. The a posteriori estimate of the error covariance is also calculated:

$$\begin{aligned} G_k &= P_k^- \cdot A^{\mathsf{T}} \cdot \left(A \cdot P_k^- \cdot A^{\mathsf{T}} + R\right)^{-1} \\ \widehat{x}_k &= \widehat{x_k^-} + G_k \cdot \left(y_k - A \cdot \widehat{x_k^-}\right) \\ P_k &= (I - G_k \cdot A) \cdot P_k^- \end{aligned} \tag{74}$$

The process of filtering the observations of travel times, denoted as tt_j, is applied to the test day d (Barceló et al., 2010), as depicted in Figure 20. It uses the predictions $\widehat{x_k^-}$ and their variances, P_k^-, calculated by the Kalman filter (73) at each step k. This helps in selecting only valid observations, denoted as OV_d^k. From the valid observations, the implemented algorithm calculates the representative observations, y_k, for the current step k by applying the statistic $EST \in \{mean, median, ...\}$ to these observations:

$$\begin{aligned} OV_d^k &= \left\{ tt_{j \in O_d^k} \;\middle|\; \widehat{x_k^-} + \alpha \cdot \sqrt{P_k^-} \;\geq\; tt_j \;\geq\; \widehat{x_k^-} - \alpha \cdot \sqrt{P_k^-} \right\} \\ y_k &= EST(\, tt_{i \in OV_d^k} \,) \end{aligned} \tag{75}$$

Here, OV_d^k is a set of observations for the test day d obtained in the time interval k.

The Kalman filter uses the values y_k to calculate from the current state $\widehat{x}_k$, based on equation (74). This updated state estimate will be used in the subsequent predictions of (74) as part of the continuous filtering process. To filter the observations, limits have been calculated. These limits are derived from the Kalman filter's prediction by adding and subtracting α times the deviation considered in the same Kalman filter, thus obtaining the upper and lower limits.

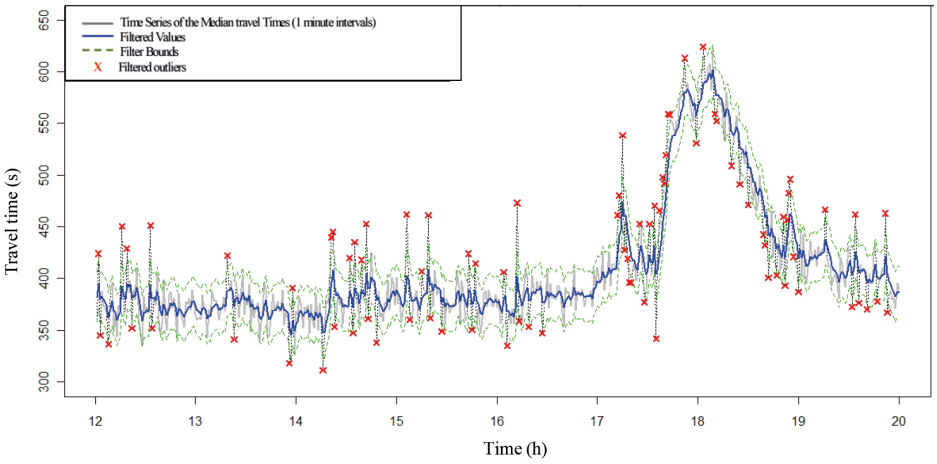

Figure 20. *An example of applying the Kalman filter to identify and remove outliers in 1-minute measurements of travel times between consecutive Bluetooth antennas on a specific test day. The outliers are then replaced with values that are consistent with observed data.*

3.1.2. Dealing with GPS data: map matching procedures

Commercial GPS data providers usually supply suitably processed data that is tailored to specific business models and applications. While the processing is a logical and natural part of their workflow, it often renders the data invalid for other general transport modeling approaches. However, unless the analyst is capable of designing their own data collection process and directly accessing the raw data, the most advantageous situation occurs when the transport analyst can obtain access to the waypoints generated by GPS. The left side of Figure 21 displays the most common and simple commercially available waypoints. Each waypoint consists of an arbitrary identity assigned by the provider to the mobile device in order to preserve the owner's identity, the date and timestamp of data collection, and the latitude and longitude corresponding to the tracked vehicle's position at that moment. The GPS data provider defines the collection policies and they can be collected at regular time intervals, after the vehicle has traveled a certain distance, at random times, and so on. The accuracy of GPS positioning can depend on various factors, such as the number of accessible satellites, signal intensity, whether the device is in an open or an urban area, and other variables. In urban areas, the accuracy is usually less than desired due to obstructions from buildings, signal interference, poor signal quality, and other factors. This can lead to erroneous positioning, as shown in the picture on the right side of Figure 21, where some waypoints are misplaced with respect to the network links, and a few may even be located on buildings.

Map-matching refers to the process of matching the geographical coordinates of waypoints to a model of the real world, such as a model of the road network in the case of tracking vehicles. The problem usually consists of relating the waypoints to the edges of the road network, which are provided by a geographic information system (GIS).

Due to its practical interest, map-matching is a widely studied problem. Kubicka et al. (2018) conducted a comparative study and application-oriented classification of selected vehicular map-matching methods covering the past two decades. The authors further provided guidelines for selecting a particular method, emphasizing that selection should be guided by the specific requirements of the application, distinguishing between offline and real-time applications. The particular case discussed in this section, as described by Cluet (2021), corresponds to a set of waypoints with high-rate positioning sampling, which will be processed using offline map-matching methods.

ID	Date	Timestamp	Latitude	Longitude
4261353	2019-11-30	07:43:58	45.445988	9.1244048
4261353	2019-11-30	07:44:11	45.445496	9.1241952
…….	……	……	……	……
4261353	2019-11-30	07:45:08	45.444767	9.1192517
4261355	2019-11-30	07:45:02	45.445980	9.1247048
4261355	2019-11-30	07:45:23	45.445574	9.1192821
…….	……	……	……	……
4261355	2019-11-30	07:46:56	45.444767	9.1197541

Figure 21. *Examples of waypoints and their misplacements.*

Geometric approaches were among the earliest used to solve the map-matching problem, due to the similarities between network link points and waypoints. A geometric map-matching algorithm uses the geometric information of the spatial road network data and primarily considers the shape of the links while disregarding their connectivity.

Given a trajectory s, geometric methods look for the most similar route in the map by using a shape similarity metric, δ. The most used similarity measures are based on distances (Hausdorff, Kim, and McLean, 2013; Fréchet, 1906), which aim to provide a good fit for the geometric aspect of the matching process.

The one-sided Hausdorff distance from curve A to curve B, as defined by Cluet (2021), is given by:

$$\delta_H^{'}(A,B) = \mathrm{Max}_{a\in A}\,\mathrm{Min}_{b\in B} d(a,b) \tag{76}$$

where $d(a,b)$ represents the Euclidean distance between points a and b. This is also known as the great circle or geodesic distance, which refers to the shortest distance between a and b on the Earth's surface. The Hausdorff distance δ_H is defined as the maximum value among the two one-sided Hausdorff distances:

$$\delta_H = \mathrm{Max}\left[\delta_H^{'}(A,B), \delta_H^{'}(B,A)\right] \tag{77}$$

As pointed out by Kubicka et al. (2018), the Hausdorff distance has some shortcomings, such as failure to account for differences between routes that use the same road segment in opposite directions. In general, any two curves occupying the same area will have a small Hausdorff distance, even when they differ significantly in shape.

A more popular distance metric is the Fréchet distance, proposed by Fréchet in his 1906 thesis.

$$\delta_F(f,g) = \inf_{\alpha,\beta} \mathrm{Max}_{t\in[0,1]} d\left\{f\left[\alpha(t)\right], g\left[\beta(t)\right]\right\} \tag{78}$$

In this equation, f, g are parametrizations of curves, $f, g : [0,1] \rightarrow R^2$, while α, β are continuous, monotone, increasing parametrization functions $\alpha, \beta[0,1] \rightarrow [0,1]$. These parametrization functions are introduced to enforce continuous and monotonically increasing parameters for f and g.

Map matching methods based on both Hausdorff and Fréchet distance metrics are sensitive to outliers in GPS positioning observations

Probabilistic approaches aim to estimate the probability that a point belongs to a single segment. One of the most common approaches is to model it in terms of a hidden Markov chain, where the transition probability represents the likelihood of a point moving from one segment to another within a given time. In urban networks with complex topology, the geometric map-matching frequently fails to provide a unique road segment solution. Therefore, combining it with a probabilistic approach, such as the Viterbi algorithm (Viterbi, 1967) is often necessary. The Viterbi algorithm is a dynamic programming algorithm for obtaining the maximum a posteriori probability estimate of the most likely sequence of hidden states that results in a sequence of observed events, specifically in the context of hidden Markov models.

For the practical purposes of modeling the road network with OpenStreetMap (OSM) and PostGIS, there are several built-in functions available for calculating the distance between a waypoint and a link, which corresponds to the initial step in the geometric approach. This OSM and PostGIS computational environment also allows us to work with pgMapMatch, an open-source Python implementation of a map-matching algorithm developed by Millard-Ball, Hampshire and Weinberger (2019).

3.2. Fusing mobile phones, GIS data, and other sources for estimating OD matrices

3.2.1. Dealing with mobile phone data

Travel demand modeling, which encompasses travel patterns and transportation mode usage, has been traditionally conducted using household techniques, as discussed in Section 1 in the summary description of the four-step model. However, the emergence of mobile smartphones with location capabilities has led to the development of novel approaches based on mobile technologies. Among them is the idea of digital diaries, which enable recording people's behaviors in urban spaces by means of probe person technology. In their seminal paper, Asakura and Hato (2004) introduce the fundamental concepts and methodologies for using smartphones to conduct tracking surveys of individuals in urban areas. While these survey techniques allow for the collection of detailed trip and traveler information, they unfortunately suffer from drawbacks, such as limitations on sample size and requiring active participation from each individual in the sample. To overcome these drawbacks, Asakura and Hato (2009) propose additional

technological developments aimed at passive data collection, later on explored by Hato (2010). Itoh and Hato (2013) suggest improvements in sampling techniques.

However, because the widespread adoption of mobile phones has made them ubiquitous and technological advancements has led to them becoming effective sensors, an alternative line of research has emerged, one which exploits *call detailed records* (CDR) of phone calls and text messages (SMS) exchanged between customers. These records are automatically collected by mobile phone service providers and offer a cost-effective and frequently updated source of data consisting of timestamps and antenna IDs. Call positions are identified according to the connected antenna, whose position is given as longitude and latitude. The time stamps can be aggregated into time intervals that align with the study's objectives. Ratti et al. (2006) were among the first to develop and utilize these techniques, and González, Hidalgo and Barabási (2008) conducted the first large-scale data evaluation of mobile phone data. Since then, this field of research has flourished in relation to the modeling and analysis of travel demand (Alexander et al., 2015; Toole et al., 2015). Moreover, the research has been applied to traffic analysis and transport models (Jiang et al., 2016; Çolak, Lima and González, 2016). These developments have progressed to the point where commercial products are now available and being exploited by companies for use in transportation projects (García-Albertos et al., 2018; Bassolas et al., 2019).

As with any other kind of observation, the huge amount of data recorded from CDR requires careful cleansing to filter out noise caused by errors in assigning mobile phones to cell towers, particularly during the tower-to-tower balancing performed by the mobile service provider. This crucial initial step is necessary to reliably extract activities and trips from CDR data. Many of the abovementioned papers dedicate specific sections to the wide variety of filtering procedures that can be applied. Subsequently, mobile phone trajectories are analyzed using data mining procedures to identify trips, represented by their start/end locations and departure times.

For such reliable inferences of activities and trips, we must distinguish between locations where users stay (where activities occur) and their moving pass-by locations (en route displacements). The conventional methods for making such distinctions are based on the agglomerative clustering algorithm proposed by Hariharan and Toyakama (2004). These methods identify points that are close in space but distant in time, along with additional criteria such as those proposed by Levinson and Kumar (1994), Schafer (2000), and Alexander et al. (2015). Essentially, these methods consist of identifying a stay point as a sequence of consecutive mobile phone records based on spatial and temporal thresholds. The spatial thresholds are set up in terms of a roaming distance, which is a parameter that must be calibrated according to network topology, phone cell density, signal quality, the accuracy of location positioning technology, and other relevant considerations.

The temporal thresholds are defined in terms of the minimum length of the stay time, which is a parameter that needs to be calibrated. This measure is calculated as the time difference between the timestamps of the first and last records at each stay

point. Users' visited locations and stay points can be categorized into types such as home, work, frequently visited places (e.g., shopping centers), and infrequently visited places. This categorization can be based on using rationale and historical evidence to record stay durations during weekdays, such as being at home between 9 pm and 8 am and at work between 9 am and 6 pm (Jiang et al., 2016). Figure 22 presents a generic situation adapted from Jiang et al. (2016), and it shows that stay points can also be clustered into stay regions. For example, in the context of defining trips based on mobile phone observations provided by CDR, let us assume that movements start from the home location in the morning and end at the home location in the evening, unless the user's distance to the home exceeds a threshold value denoted as d_{Max}. A threshold value d_{Min} defines a minimum movement distance to identify successive records.

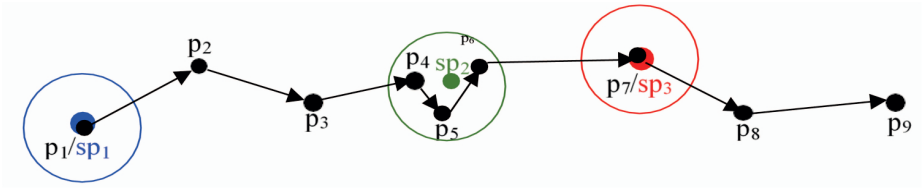

Figure 22. *p_i represents the i-th observation of a mobile phone, and st_j, denotes the j-th stay-point (home sp_1, shopping center sp_2, work place $sp_3 \ldots$). Mobile records p_4, p_5, and p_6 are clustered into stay point sp_2. Circles identify the thresholds.*

Algorithm 1:
main()
1 for each user *u*
2 for each day *a*
3 for each CDR observation *k*
4 let p_{uak} = position for event *k*
5 if(trip_active == false)
6 trip_active = detect_trip_start()
7 end
8 if(trip_active == true)
9 trip_ended = detect_trip_end()
10 end
11 if(trip_ended)
12 store_trip()
13 end
14 end
15 end
16 end

Algorithm 2
detect_trip_start()
17 if (trip_set empty)
18 if(p_{uak} != homebase and d(p_{uak},homebase) < d_{max} and d(p_{uak},homebase) > d_{min})
19 trip_active = true
20 origin = homebase
21 end
22 if(p_{uak} != homebase and d(p_{uak},homebase) > d_{max})
23 trip_active = true
24 origin = p_{uak}
25 end
26 if(p_{uak} == workbase and d(p_{uak},homebase) > d_{max})
27 trip_active = true
28 origin = homebase
29 destination = workbase
30 end
31 else
32 if(p_{uak} != previous_trip_start(trip_set) and d(previous_trip_start(trip_set), p_{uak}) > d_{min})
33 origin = previous_trip_start(trip_set)
34 end
35 end

Algorithm 3
detect_trip_end ()
36 if(p_{uak} == workbase or p_{uak} == homebase)
37 destination = p_{uak}
38 else
39 if($p_{ua(k+1)}$ exists)
40 if(p_{uak} == $p_{ua(k-1)}$)
41 destination = p_{uak}
42 end
43 else
44 if(d(p_{uak},homebase) < d_{max})
45 destination = homebase
46 else
47 destination = p_{uak}
48 end
49 end
50 end

Exhibit 1. Trip Generation Algorithms

The trip generation algorithm (Gundelgård et al., 2015, 2016) shown in Exhibit 1 uses CDR observations and consists of three functions. The main function (Algorithm 1) is called *main*() and loops through all available CDR observations for each user and for each day. It scans each observation, invoking the *detect_trip_start*() function (Algorithm 2) to determine whether the trip start condition is met. This condition is satisfied if the distance from the ending point of the previous trip (line 33 of the pseudocode) or from the home position for the first trip of the day (line 19) exceeds the value d_{Max}. While a trip is in progress, the *detect_trip_end*() function (Algorithms 3) is invoked for every observation. The algorithm considers a trip has ended if the user arrives at home (line

37), work (line 45), or if two consecutive events have the same position (line 41). When a trip has ended, the *main()* function repeats the process and tries to detect the user's next trip start by calling *detect_trip_start()*.

Additionally, it is possible to extract stay locations with durations exceeding a certain threshold (a parameter defined by the analyst, depending on the context). If the observation period is sufficiently long, all frequently visited stay locations can be identified. Trips can also be filtered and aggregated into time intervals to estimate dynamic OD matrices. During these periods, origins and destinations can be estimated by identifying the most common positions and associating them with TAZs, particularly when the origin zone differs from the destination zone. Identifying the area of study presents another significant consideration when generating OD matrices from mobile phone data. Urban databases typically store socioeconomic, population, and other related data at the level of census tracts. The travel patterns defined by OD matrices correspond to traffic analysis zones (TAZ), as discussed in Section 1. Determining how to define TAZ splitting is not a trivial task and usually takes into account the socioeconomic characteristics of the studied region, which aligns with considerations about the underlying causes of mobility (Ortuzar Willunsen, 2011). TAZs are usually well-balanced in terms of population and demand analysis criteria, and they are frequently formed by aggregating census tracts through a clustering process that covers the entire territory. Finally, the cellular cells associated with each mobile phone antenna form the third layer covering the territory, which may not have a direct correspondence with the other two partitions covering the territory.

Consequently, in order to avoid significant errors caused by misalignments and inconsistencies between the three coverings, careful design is necessary (Zhang et al., 2010; Iqbal et al., 2014; Montero et al., 2019). For example, Bassolas et al. (2019) propose a heuristic to overcome the lack of exact correspondence between Voronoi cells. Their method assigns residents located in a given Voronoi cell to one of the intersecting census tracts or neighboring areas. The assignment probability is directly proportional to the square of the population of the census tract and inversely proportional to the square of the number of users already assigned to that tract. This assignment process ensures a local homogeneous sample density among neighboring census tracts. Figure 23 (adapted from Gundelgård et al., 2015, 2016) depicts an example of Voronoi tessellation modeling the phone cells and assignments of trips to TAZs. These TAZs are the result of aggregating Voronoi polygons based on mobile phone data for Senegal, which was obtained from the mobile operator Orange (de Montjoye et al., 2014). The data comprises call detail records (CDR) of phone calls and text exchanges (SMSs) between customers in Senegal, collected between January 1, 2013, and December 31, 2013.

The data used in Gundelgård et al. (2015, 2016) consists of 1666 antenna IDs and their corresponding locations, as well as mobility data for a year. The mobility data is based on a rolling two-week basis and comprises approximately 300,000 randomly sampled users.

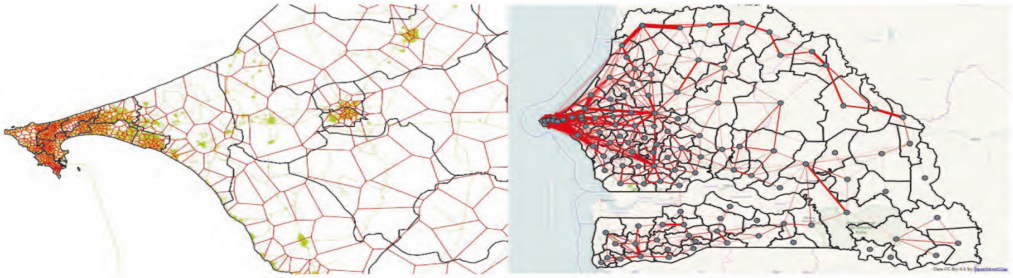

Figure 23. *Example of Voronoi tessellation and trip assignments to TAZ.*

Figure 24 summarizes the described methodological process. The upper part of the logic diagram represents the processing of the CDR. It is important to note that the proposed process produces a global OD matrix, which includes all trips regardless of the transportation mode used. In other words, there is no distinction between trips by car, bus, or any other public transport mode such as metro or railway. However, as discussed in Section 2 when describing the inputs to various models (particularly DTA and microscopic models), the required input is a dynamic time-dependent OD matrix $X(t) = [X_{rs}(t)]$ specific to each transport mode. For example, in the practical cases addressed in this paper, separate OD matrices are needed for car trips. Therefore, an additional step is needed to generate such OD matrices.

The most common solution to this problem is depicted in the conceptual diagram in the lower part of Figure 24. It consists of integrating the OD^{CDR}, which was initially estimated from the CDR processing, with other data sources that explicitly account for modal splitting. The most typical source is the household transport survey, which has long been used in transport demand analysis. Household surveys have the disadvantage of representing a small sample of the whole population and providing a kind of snapshot that is valid only for the time when the survey was conducted. However, they offer the advantage of being produced by carefully designed samples using well-established statistical sampling techniques that ensure being able to reliably extend the sample to the whole population. The fusion of this (possibly outdated) historical OD^{H} with the more accurate and updated OD^{CDR} is then used to derive a set of mode-specific matrices (OD^{mode}). This can be achieved by establishing correspondences between the splitting rates of the initial OD and assuming that they will prevail in the second (Montero et al., 2019). Alternatively, historical data can be used to calibrate a discrete choice model (as discussed in Section 2.1) and apply it to the OD^{CDR} to estimate the modal ODs.

Once a modal OD has been obtained, such as the OD^{car} for car trips, it can be refined if additional data sources are available. The most usual case is when conventional traffic data, link flows, and speeds, are accessible available from the traffic management system operating in the corresponding area. The estimation of OD matrices from available traffic measurements to generate inputs to transport models is a notably problem that has garnered substantial attention from researchers. This attention has been driven by its relevance for practical applications, especially in recent years with the growing

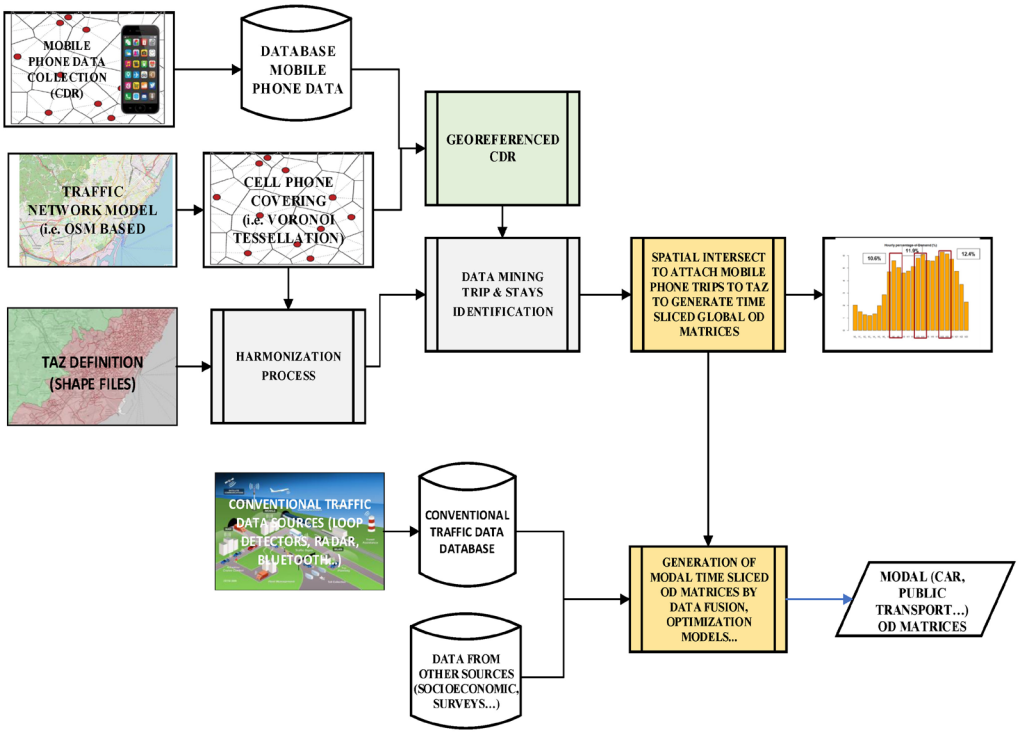

Figure 24. *Methodological diagram of the procedures for generating OD matrices from CDR.*

demand for dynamic models that require dynamic inputs (see Antoniou et al., 2016 for an overview).

The basic assumption is that, given an OD matrix X, an equilibrium assignment (as described in Sections 2.1 and 2.2) provides estimates of the link flows Y and, in some cases (e.g., dynamic assignments), estimates of other traffic variables such as paths, partial paths, and travel times. The reciprocal problem, as discussed by Cascetta (2001), can be formulated as follows: Assuming that $\widehat{Y}$ is a set of observed link flows in a subset of links in the network, the goal is to find the OD matrix $\widehat{X}$ whose equilibrium assignment onto the network will generate the observed flows. Mathematically, the problem is highly underdetermined and challenging, but acceptable solutions can be found by imposing additional constraints (Cascetta, 2001; Antoniou et al., 2016). It can be formulated as a nonlinear optimization problem in various forms, such as the following highly suitable bilevel optimization problem:

$$\begin{aligned} &\text{Min } Z(X) = \mathcal{F}\left(X, \widehat{X}, Y, \widehat{Y}\right) \\ &s.t.\ \widehat{Y} = \text{assignment}\left(\widehat{X}\right), \quad \widehat{X} \in \Omega \end{aligned} \tag{79}$$

where $\mathcal{F}$ is typically a distance function that measures the difference between a target or historical matrix X and the estimated matrix $\widehat{X}$, as well as between the observed link flows Y and the estimated flows $\widehat{Y}$. The feasibility dominion Ω is usually determined by additional constraints (Djukic, 2014; Antoniou et al., 2016). To enhance computational efficiency, the most efficient linearization approaches are often employed to approximate the assignment used in (79). This is especially true in dynamic cases, when the OD matrix is time-dependent and replaces the assignment mapping with:

$$y_{lt} = \sum_{ijr} a_{ijr}^{lt} x_{ijr} \;\rightarrow\; Y = A(X)X \tag{80}$$

Here, a_{ijr}^{lt} is known as the assignment matrix and represents the proportion of the OD flow departing from origin i at time r and going to destination j, crossing link l at time $t \geq r$. Various linearization approaches have been proposed in practice, such as those presented by Toledo and Kolechkina (2013), Frederix, Viti and Tampère (2013), and Ros-Roca et al. (2022). Several algorithmic approaches have been proposed to solve the model, such as those by Toledo et al. (2013), Antoniou et al. (2016), and Ros-Roca et al. (2022). There are variants of the simultaneous perturbation stochastic approximation (SPSA) method originally proposed by Spall (1992) that are explored in Antoniou et al. (2016) and Ros-Roca, Montero and Barceló (2020), among others. Another noteworthy approach is simulation-based optimization, as described by Osorio (2019), which is well suited to dynamic cases involving a simulation-based approach.

3.2.2. OD Estimation and GPS data

An emerging trend enabled by the accessibility of GPS traces is the development of so-called "data-driven" models in which the parameters of the mathematical model of the transport system are directly estimated from ICT measurements. These approaches rely on large samples of vehicle data collected over a sufficiently long period. The first step in the process, as discussed in Section 3.1.2, consists of obtaining individual vehicle trajectories from GPS records and map-matching these waypoints onto the graph of the transportation network by means of specialized map-matching algorithms suitable for the available sample, whether it has low or high sampling rates. Assuming that the first record corresponds to the start of the trip and the last to its end, and considering the information recorded in the waypoints (date, time tag, longitude, and latitude), then the corresponding trajectory can be associated with a specific departure zone and destination zone for a given day and time.

This zone assignment yields a primary set of OD matrices for each day and time, although these OD matrices correspond to segments of the total population and are strongly biased, since they represent only users of the GPS technology utilized to collect the data. Thus, the sample is not necessarily representative and there are no clear methods for expanding it to the whole population. However, the identified paths in the network can be clustered and analyzed using techniques such as machine learning techniques (Lopez et al., 2017a, 2017b) to identify the paths used and the proportion of their

usage for each OD pair. Path choice models like those in (69) can empirically estimate their parameters from the recorded data (Krishnakumari et al., 2019; Ros-Roca et al., 2022). Furthermore, when an equipped vehicle crosses a link and generates a waypoint and correctly map-matches it to the corresponding link, the processed data provides information such as link identity, the time of crossing the link, the origin of the trip, the departure time (time-tag of the first waypoint recorded), and the trip destination. With this information, it is possible to estimate an empirical assignment matrix that allows a reformulation of (80) in order to relate the estimated traffic count to the OD flows:

$$y_{\ell t} = \sum_{(i,j)\in N} \sum_{r=1}^{t} \bar{a}_{ijr}^{\ell t} x_{ijr} \tag{81}$$

where y_{lt} represents the estimated flow in link l at time t; x_{ijr} is the flow departing origin i to destination j at time interval $t \in T$; and $\bar{a}_{ijr}^{lt}$ is the estimated assignment matrix, which is the fraction of trips from origin i to destination j, departing from i at time interval r reaching link l at time t, estimated from the GPS traces. In other words, this data-driven approach reformulates the OD calibration problem by replacing the analytical approaches for estimating the model parameters with empirical estimations from ICT applications. Behara (2019) proposes an alternative approach based on estimating partial path travel times from Bluetooth measurements obtained suitably located antennas in the network.

Nevertheless, all these approaches by Krishnakumari et al. (2019), Mitra et al. (2020), Behara (2019), and Ros-Roca et al. (2022) still complete the model formulation in (79) by optimizing the value of an objective function $\mathcal{F}\left(X,\widehat{X},Y,\widehat{Y}\right)$ that minimizes a distance measure between the estimated $\widehat{X}$ of the OD matrix and a target OD matrix X, as well as between the estimated link flow counts $\widehat{Y}$ (obtained from (81)) and the observed link flow counts Y (or the estimated and measured partial path travel times).

3.3. Extracting traffic data from image processing

In 2001 the Federal Highway Administration (FHWA) initiated an intense debate about the validity and application of traffic simulation models for traffic analysis. Consequently, in 2002, the FHWA Traffic Analysis Tool Program (https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm) was launched to address the questions raised and to establish a methodological framework for the construction and utilization of transport models. The FHWA acknowledged (Alexiadis, Colyar and Halkias, 2007) that microscopic traffic simulators can help evaluate complex scenarios by intricately modeling real-world transportation networks, a task that is challenging using more conventional methods. Moreover, advancements in computer technologies have enabled these simulators to model larger and more complex transportation systems, thereby supporting associated decision-making processes.

From the very beginning, the stakeholders involved in the program unanimously agreed on the premise put forth in this paper: Microscopic models need data, particularly detailed microscopic data that are not easily obtained and not always available. A

comprehensive understanding of microscopic traffic flow and car following behavior is crucial for advancing traffic flow theory. This understanding is essential for constructing traffic simulation models, and the most effective means of acquiring such knowledge is by collecting empirical data and providing it as evidence. Consequently, a companion program called the *next generation simulation* (NGSIM) program (https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm) was launched with the aim of developing driver behavior algorithms for microscopic modeling by collecting detailed, high-quality traffic datasets. Multiple data collection sites were equipped, and the collected datasets are freely available at the corresponding websites. Notable among them are Interstate I-80 Highway (https://www.fhwa.dot.gov/publications/research/operations/06137/index.cfm) and US Highway 101 (https://www.fhwa.dot.gov/publications/research/operations/07030/index.cfm).

The I-80 Highway location is shown in Figure 25. According to the dataset provided in the website's NGSIM Fact Sheet, "the NGSIM program collected detailed vehicle trajectory data on eastbound I–80 in the San Francisco Bay area in Emeryville, CA, on April 13, 2005. The study area was approximately 500 meters (1,640 feet) in length and consisted of six freeway lanes, including a *high-occupancy vehicle* (HOV) lane. An onramp also was located within the study area. Seven synchronized digital video cameras, mounted from the top of a 30-story building adjacent to the freeway, recorded vehicles passing through the study area. NG-VIDEO, a customized software application developed for the NGSIM program, transcribed the vehicle trajectory data from the video. This vehicle trajectory data provided the precise location of each vehicle within the study area every one-tenth of a second, resulting in detailed lane positions and locations relative to other vehicles. A total of 45 minutes of data are available in the full dataset, segmented into three 15-minute periods: 4:00 p.m. to 4:15 p.m.; 5:00 p.m. to 5:15 p.m.; and 5:15 p.m. to 5:30 p.m. These periods represent the buildup of congestion, or the transition between uncongested and congested conditions, and full congestion during the peak period. In addition to the vehicle trajectory data, the I–80 dataset also contains computer-aided design and geographic information system files, aerial orthorectified photos, freeway loop detector data within and surrounding the study area, raw and processed video, signal timing settings on adjacent arterial roads, traffic sign information and locations, weather data, and aggregate data analysis reports".

Video image processing for traffic analysis remains more of an art than a science. While automated tools can provide an initial approximation, it is no easy task to achieve the level of precision required for extracting sufficiently accurate empirical vehicle trajectories to develop traffic flow models. Because even the best image processing tools cannot overcome the inherent complexities of projection errors, occlusions, shadows, the non-rectilinear shapes of real vehicles, and vehicles with colors similar to the pavement, significant human intervention is still required if traffic flow theory is to advance. Figure 26 depicts the propagation of shockwaves collected from the trajectories at the US101 site.

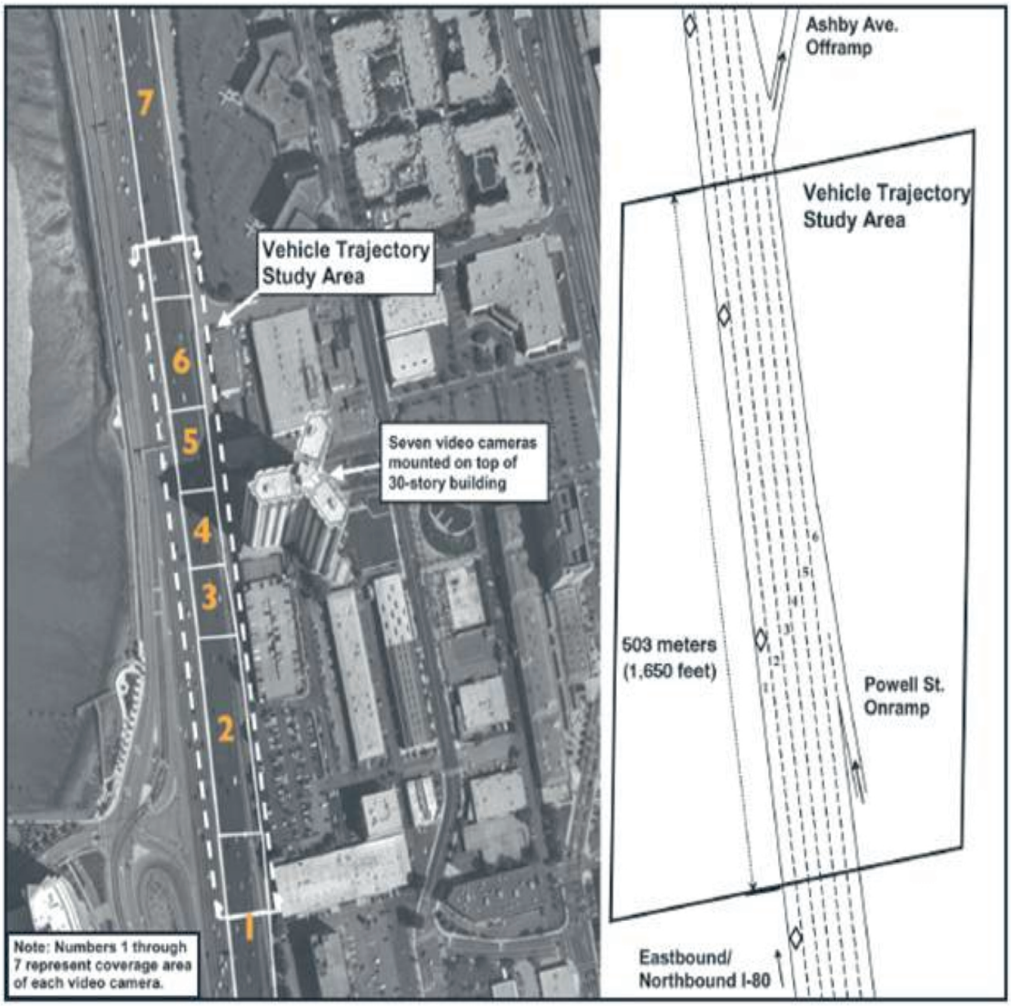

Figure 25. *The aerial photograph on the left shows the extent of the I–80 study area relative to the building where the video cameras were mounted, along with the coverage area for each of the seven video cameras. The schematic drawing on the right shows the number of lanes and location of the Powell Street onramp within the I–80 study area. Source: Public Domain "Federal Highway Administration Research and Technology" https://www.fhwa.dot.gov/publications/research/operations/06137/index.cfm.*

The detailed microscopic data collected by NGSIM were expected to serve as valuable resources for validating traffic simulation models by comparing the values of various vehicle kinematic variables, which include the time and space headways that could be measured, speed distributions, accelerations, and changes in acceleration (jerks). By analyzing these data, we anticipated being able to estimate the parameters of the car-following and lane change models, enabling them to accurately reproduce the observed values.

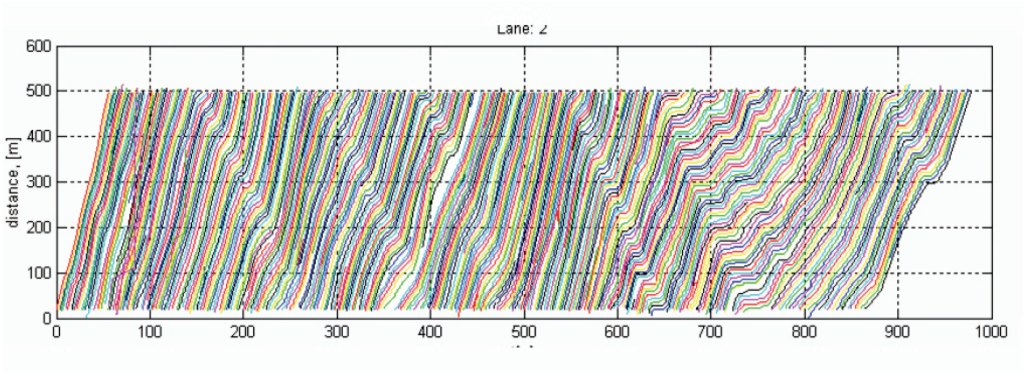

Figure 26. *Shockwave Characteristics (NGSIM, I-80 Dataset, 5:00–5:15 pm, Lane 2).*

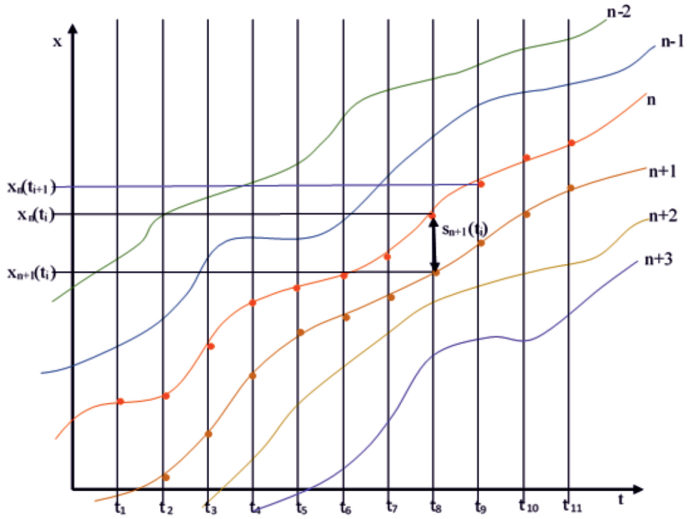

Figure 27. *Hypothetical reconstruction of vehicle trajectories from video processing.*

To summarize the process of extracting information from the trajectories using a hypothetical example, Figure 27 represents a few trajectories obtained through suitable processing of the video data recorded at constant time intervals $\Delta t = t_{i+1} - t_i$ ($\Delta t = 0.1$ seconds for the NGSIM data). The red trajectory corresponds to the *n-th* vehicle, and the red dots along the trajectory indicate the vehicle's positions at each instant t_i, $i = 1, \ldots, T$.

For a given instant t_i, the relative positions of two consecutive vehicles, a leader n and its follower $n+1$, are denoted by $x_n(t_i)$ and $x_{n+1}(t_i)$, respectively. The space headway between them can be calculated as:

$$s_{n+1}(t_i) = x_n(t_i) - x_{n+1}(t_i) \tag{82}$$

For each trajectory, the corresponding time series of speeds can be calculated. For instance, Coifman and Li (2017) propose using the mean difference over multiple time intervals as follows:

$$\widehat{v}_n(t_i) = \frac{x_n(t_{i+p}) - x_n(t_{i-p})}{2p\Delta t} \tag{83}$$

Similarly, accelerations can be derived:

$$\widehat{a}_n(t_i) = \frac{\widehat{v}_n(t_i) - \widehat{v}_n(t_{i-1})}{2\Delta t} \tag{84}$$

Additionally, the jerks (the time change of the acceleration) can be calculated. These and other values can be used to calibrate the parameters of car-following models and test their quality. One early example can be found in Yeo and Skabardonis (2007), who develop, calibrate, and test an improved car-following model based on empirical observations of NGSIM trajectories. Another example is Bevrani and Chung (2011), who modified the Gipps car-following model to enhance its capabilities for safety studies. In the first case, they use the analysis of the trajectories to estimate the probability distribution of space headways estimated from equation (1) and the speed distribution from equation (2). In the second case, the study primarily focuses on the enhanced car-following model's ability to predict the expected time to collision (TTC), a critical indicator of a given traffic situation. TTC for the follower vehicle *n + 1* is calculated as:

$$TTC_{n+1}(t_i) = \frac{x_n(t_i) - x_{n+1}(t_i) - l_n}{\widehat{v}_{n+1}(t_i) - \widehat{v}_n(t_i)} \tag{85}$$

where $\widehat{v}_{n+1}(t_i) > \widehat{v}_n(t_i)$.

The analysis of the I-80 NGSIM data conducted by Yeo and Skabardonis (2007) revealed that the probability distribution of space headways under congested conditions follows a lognormal distribution. The mean of the distribution was found to be 4.24 meters, with a variance of 14.6035 (see Figure 28). A similar distribution was found for the shockwave speeds, and Bevrani and Chung (2011) also found similar lognormal distributions for the TTC.

The theoretical expressions for space headways (82), speeds (83), accelerations (84), and TTC (85), as well as other derived estimates such as jerks or shockwave speeds, are used to estimate empirical values that are later employed for the calibration and validation of car-following models. These expressions implicitly assume that either the empirical values are error-free or their errors have been minimized. However, this assumption is unfortunately not always met. As already discussed in this paper, errors can affect the observed points regardless of the technique used to collect vehicle positions. In the case of the trajectories recorded after processing the video images, these points may be dispersed in the vicinity of the actual physical path followed by the vehicle. These measuring errors can substantially impact the analysis of a follower's and leader's consecutive vehicle behaviors (Punzo, Borzacchiello and Ciuffo, 2011). If $\widehat{f}_n(t)$ is the trajectory function of vehicle n, measurement errors introduce noise into $\widehat{f}_n(t)$,

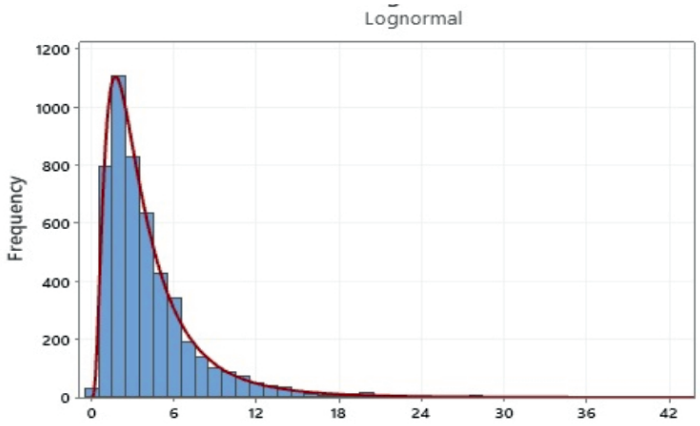

Figure 28. *Lognormal probability distributions of space headways for I80.*

which is magnified by differentiation when calculating speeds, accelerations, and jerks. These quantities represent the physical variations in acceleration over time and can be considered as random components in the traveled space.

Direct use of raw data reveals unacceptable accelerations and physically unreliable jerks. For example, analyzing the acceleration distribution over the entire datasets reveals unfeasible extreme values and anomalous shapes in the distributions, which Punzo et al. (2011) interpret as clear indications of problems in the data collection. They also prove that their analytical evidence shows a positive bias in $\widehat{f_n}(t)$, due to a systematic error component that they believe is inherent to this type of measurement. However, this bias is not self-evident when looking at the trajectory of a single vehicle but becomes apparent when examining the trajectories of consecutive vehicle pairs. Their paper provides analytical evidence of the bias propagated in the vehicle trajectory functions, for which they propose consistency requirements. In a later paper, Montanino and Punzo (2015) refine the procedures for properly reconstructing trajectories, minimizing errors, and making the data useful for the intended purposes. They propose a systematic error analysis based on geometrical and physical considerations, combined with filtering and smoothing techniques. Specifically, they apply Gaussian kernel smoothing to the position data for each trajectory to reduce the impact of noise resulting from data reduction.

Lu and Skabardonis (2007), in a companion paper to Yeo and Skabardonis (2007), also identify the disturbances in the trajectories from NGSIM data due to measurement noise that must be corrected before being used for traffic studies. They apply a Butterworth low pass filter, as proposed by Butterworth (1930).

Many other researchers have identified these limitations of NGSIM datasets and proposed various error filtering and data smoothing techniques to correct them. However, other researchers such as Coifman and Li (2017) claim that "the NGSIM errors are beyond anything that could be corrected strictly through cleaning or interpolation of the reported NGSIM data." In analyzing NGSIM trajectories, they found that their tracking of vehicles quite frequently results in vehicular collisions. Consequently, they system-

atically remove the NGSIM trajectories to generate a subset of different trajectories that are free of blatant errors. They also employ similar filtering and smoothing techniques, particularly those that use the Savitzky-Golaydigital filter (Savitzky and Golay, 1964). The techniques aim to enhance the precision of the data without distorting its underlying trends, and they involve using a convolutional process that fits successive subsets of datasets in a way similar to a sliding window with a low-degree polynomial. Other authors employ spline fitting methods, resulting in a new dataset that Coifman (2017) has made publicly available for research purposes. He asserts that these cleaned NGSIM data can now serve as a benchmark for assessing the quality of trajectory data.

4. Concluding remarks and insights into some current trends

The main thesis held in this paper is that a proper understanding of a complex system can be achieved by acquiring adequate knowledge about the system and translating it into a modeling hypothesis. The hypothesis should serve as the foundation for explaining how the system works and formally representing it through a model. Models, therefore, become a scientific tool for better understanding a system and supporting rational decisions by providing insights into how the system will behave under other conditions. In other words, models provide answers to what-if questions about the system. As emphasized in Section 1, the system, the observer, and the model form a unit known as the Minsky triad, in which questions are asked about the system and its objectives in order to support rational decision-making. Thus, no unique model of a system exists but instead multiple models that depend on the specific questions asked by the observer. This general modeling theory, outlined in Section 1, applies to various types of systems and specifically to transport systems, which is the focus of this paper.

Transport systems belong to a family of complex systems that can be analyzed using the Minsky triad. Section 2 provides brief examples of the three main families of transport models based on the modeler's perspective regarding the questions that the models aim to answer and the corresponding modeling hypotheses that align with the characteristics of the system that are relevant for answering these questions. Each modeling approach, whether macro (static or dynamic), micro, or meso, is summarized in a subsection that describes the underlying modeling hypotheses and how they are translated into a mathematical formal representation. The models resulting from each approach identify the parameters on which they depend, and the numerical values these parameters must be estimated based on the data.

Another key thesis of this paper holds that data required by models are not in themselves information but instead carry information that requires specific processing. This establishes a two-way interaction between models and data. Models need data, and data can provide only limited useful information without the aid of models. Models are essential for bridging the gap between descriptive and predictive capabilities, as they provide an understanding of the system.

For models to be truly valuable, calibration and validation are necessary. This entails ensuring that the model's parameter values are accurate, thereby establishing the validity of the model and its ability to faithfully reproduce the system's behavior. Section 2 concludes by providing an overview of the calibration and validation processes, emphasizing the pivotal role of data in these processes.

Considering the significant role of data in modeling approaches and the fact that data alone do not inherently reveal their embedded information, it becomes imperative to address critical issues pertaining to data availability, its characteristics based on the employed data collection technology, and the appropriate data processing techniques required to extract the relevant information about each phenomenon generated by this data. This information is crucial for parameter estimation during the calibration process, as well as for comparison during the validation process, ultimately enabling the utilization of the model to answer what-if questions.

Section 3 addresses these topics be establishing a methodological framework for data processing. It provides a general overview of the various types of data and their characteristics, depending on the available technologies. The section also illustrates the use of this methodological framework with a few selected examples based on some of the most recent data collection technologies, namely those supported by ICT applications like Bluetooth, CDR and GPS from mobile devices, and video image processing.

However, this section begins by emphasizing that datasets, regardless of the technology used, always contain errors, missing data, and other flaws that need to be corrected and completed to ensure data completeness and consistency. This is achieved through the application of filtering techniques, one example of which is the powerful Kalman filter, which measures travel times by tagging two consecutive antennas used by mobile Bluetooth devices. The Kalman filter not only identifies and removes outliers, but it also replaces them with the most likely values to obtain a complete and consistent dataset.

The subsequent steps demonstrate the utilization of data provided by two prevalent ICT applications to generate dynamic OD matrices, which serve as crucial inputs for microscopic and mesoscopic traffic models. Dynamic OD matrices reveal the time-dependent traffic patterns, which can be identified by techniques that either track the CDRs of mobile phones associated with phone cells corresponding to the antennas along their paths or record the GPS waypoints that track the trajectory of mobile devices. In both cases, ad hoc filtering procedures are necessary to remove erroneous records or ensure the validity of the records. This involves ensuring correct matching between CDR and geographic coverage of TAZs and phone cells for mobile phones, as well as map matching between waypoints and their physical locations on the road network. These specific filtering processes are briefly discussed and illustrated. However, the OD estimates in both cases have limitations. They may either provide global estimates of trips without distinguishing the mode of transport used, or they only correspond to a specific mode, such as cars, for which only a subsample is available (i.e., equipped vehicles). Consequently, additional information from other sources is required in both scenarios, either to split the OD into the various transport modes or to find a way to extend the

sample to the whole population. The most commonly used sources are prior information from conventional surveys in the form of target OD matrices and link flow counts from conventional detection stations that serve as reference ground truth. This section of the paper indicates how specific optimization techniques can be used to achieve the objectives.

Finally, Section 3 concludes with a representative example of using video image processing to extract traffic information, specifically from the FHWA's NGSIM project, aimed at providing traffic datasets for testing traffic simulation models, particularly car-following models that are fundamental to microscopic simulation engines. The section describes the datasets, their processing, and the filtering techniques employed, while also highlighting the controversy surrounding the datasets since their inception. This example effectively demonstrates the advantages and disadvantages associated with certain uses of technologies in extracting valid data and how these challenges have been overcome.

We could conclude here, as these remarks have highlighted how the thesis stated at the beginning of the paper has been demonstrated through significant examples. By identifying the hypotheses underlying the key transport models and illustrating the interdependence between models and data, it is evident that each relies on the other and neither can replace or render the other unnecessary. However, ending at this point may leave a sense of incompleteness. It is important to provide a glimpse of current trends and what lies ahead.

Numerous avenues of exploration can be identified, but two dominant themes emerge, considering that models used to analyze transport systems are also tools for analyzing the mobility they facilitate. These themes seek to offer insights into the question: What factors can determine the urban mobility of the future?

4.1. Scenarios dominated by technological developments: the case of connected and autonomous vehicles

For those who believe that the future of mobility will be fundamentally determined by technology such as connected and autonomous vehicles (CAV), electric vehicles, and information and communication technologies (ICT), understanding the future of mobility requires models that take into account the influence of these technologies. This perspective implicitly assumes that technology will enable people to travel from origins to destinations for the same reasons as today while selecting the most convenient paths, but with the advantage of CAVs making choices based on data collected from other CAVs in addition to conventional information. The key modeling challenge then becomes how car-following models will function, not only for pairs of vehicles but also for groups or platoons of interconnected vehicles traveling in a coordinated manner. It is crucial to determine the conditions under which the dynamics of the platoon will remain stable. To provide a comprehensive overview, it is worth mentioning a seminal work inspired by the car-following approaches discussed in Section 2. Building upon the general modeling approach proposed by Ward and Wilson (2011) and Wilson (2011), which formulates car-following models in terms of a functional framework modeling the follower's re-

action in terms of acceleration or deceleration based on speeds, spacings, and relative speeds:

$$a_{n+1}(t) = \ddot{x}_{n+1}(t) = \mathcal{F}[s_{n+1}(t), \Delta v_{n+1}(t), v_{n+1}(t)] \tag{57}$$

These models have "uniform flow" steady solutions (equilibria) if, for each $s^* > l$, there is a $v^* = V(s^*) > 0$ such that $\mathcal{F}(s^*, 0, v^*) = 0$, where $V(s^*)$ is the equilibrium speed-spacing relationship that leads to a fundamental diagram. Researchers such as Wagner (2016) and Talebpour and Mahmassani (2016) deem this functional framework to be a suitable starting point for studying the behavior of autonomous vehicles, due to its generic approach that does not assume specific driver characteristics and can therefore capture interactions among autonomous vehicles with nonhuman drivers.

Ward and Wilson (2011) define the string stability of a platform in terms of the response to a leader suddenly braking and the decaying perturbation as it propagates upstream within the platoon. In this case, the car-following model is considered platoon stable. Then the analytical conditions string stability can be expressed in terms of the partial derivatives $\mathcal{F}_s$, $\mathcal{F}_{\Delta v}$, $\mathcal{F}_v$ of the functional $\mathcal{F}[s(t), \Delta v(t), v(t)]$, evaluated at $(s^*, 0, v(s^*))$, as follows:

$$\begin{aligned}
\mathcal{F}_s^n &= \left.\frac{\partial \mathcal{F}(s_n, \Delta v_n, v_n)}{\partial s_n}\right|_{(s^*,0,v(s^*))} \\
\mathcal{F}_{\Delta v}^n &= \left.\frac{\partial \mathcal{F}(s_n, \Delta v_n, v_n)}{\partial \Delta v_n}\right|_{(s^*,0,v(s^*))} \\
\mathcal{F}_v^n &= \left.\frac{\partial \mathcal{F}(s_n, \Delta v_n, v_n)}{\partial v_n}\right|_{(s^*,0,v(s^*))}
\end{aligned} \tag{85}$$

and:

$$\sum_n \left[\frac{\mathcal{F}_v^{n2}}{2} - \mathcal{F}_{\Delta v}^n \mathcal{F}_v^n - \mathcal{F}_s^n\right] \left[\prod_{m=n} \mathcal{F}_s^m\right]^2 \tag{86}$$

where the index n covers the set of vehicles in the platoon. In the case of Talebpour and Mahmassani (2016), string stability is evaluated in terms of the intelligent driver model (IDM) developed Kesting et al. (2010) and defined by equations (55) and (56), where each vehicle in the platoon will have specific model parameters associated with it.

Considering that the autonomous vehicles are equipped with monitoring capabilities for all vehicles in their vicinity, their time lags and anticipation times can be estimated in terms of sensing and mechanical delays. The speeds of autonomous vehicles in the platoon should allow them to come to a full stop when the leader initiates maximum deceleration by braking.

They analyze the string stability of the proposed model following the approaches of Ward (2009), Ward and Wilson (2011), and Treiber and Kesting (2013) for a homogeneous platoon of vehicles. The partial derivatives (86) are calculated to evaluate it in

terms of equation (87).

$$\begin{aligned}
\mathcal{F}_s &= \frac{2a}{s^*}\left(\frac{s_0 + Tv(s^*)}{s^*}\right)^2 \\
\mathcal{F}_{\Delta v} &= -\frac{v(s^*)}{s^*}\sqrt{\frac{a}{b}}\left(\frac{s_0 + Tv(s^*)}{s^*}\right) \\
\mathcal{F}_v &= -\frac{a\delta}{v_0}\left(\frac{v}{v_0}\right)^{\delta-1} - \left(\frac{2aT}{s^*}\right)\left(\frac{s_0 + Tv(s^*)}{s^*}\right)
\end{aligned} \tag{87}$$

The resulting partial derivatives are expressed as functions of the vehicle speed v, the equilibrium gap s^*, and the equilibrium speed $v(s^*)$. These expressions can be simplified using the equilibrium relationships proposed by Treiber and Kesting (2013):

$$s^*(v) = \frac{s_0 + vT}{\sqrt{1 - \left(\frac{v}{v_0}\right)^{\delta}}} \tag{88}$$

From this, Talebpour and Mahmassani (2016) conduct an analysis of stable and unstable scenarios based on the parameter values governing the model, taking into account the suggested values from empirical evidence of cruise control studies. This example is selected to be consistent with the models discussed in Section 2, which can only be studied analytically or through simulation, since the real systems are not yet implemented. It illustrates how models can assist in the design and testing of new systems. Furthermore, considering that time-lags can mainly depend on sensing delays, which are strongly influenced by telecommunication technologies, modeling approaches that also include telecommunication aspects have been explored. One early example is Talebpour, Mahmassani and Bustamante (2016), and a more recent one is Dai et al. (2022). Let us close these comments by mentioning other types of modeling approaches to car following models, such as those inspired by reinforced learning processes (Wu et al., 2017). So far, these approaches can only be tested through simulations due to the lack of observed data.

4.2. Scenarios dominated by other factors: urban forms, accessibility, etc.

For those who acknowledge the influence of non-technological factors, such as urban forms and their impact on the temporal and spatial distribution of activities, the future of mobility is intertwined with the evolution of cities and the complex relationships between mobility, urban forms, and transport systems. This perspective assumes that technology enables new possibilities like telecommuting, which eliminates the need for physical displacement to overcome physical distances, or the concept of the ”15-minute city,” where urban areas are designed to reduce the necessity of extensive travel by prioritizing non-motorized modes of transportation over motorized ones. Consequently, alternative models are required to explore these aspects. What are these models?

We need to shift our mindset, as suggested by Barceló (2019), from a conventional reductionist approach to a complex dynamic systems approach. In this perspective, a complex system is a system composed of a large number of interacting components that, as a whole, exhibit properties that are distinct from the properties of the individual components. This implies taking a holistic view of the whole as different from the sum of the parts. Thus, the transport system comprises various interconnected networks for different modes of transportation, such as cars, buses, metro, and railways, which need to be integrated to accurately capture their interactions. This paradigm shift challenges the traditional modeling approach focused solely on trips and their purposes, and instead seeks to understand the underlying causes and consequences. Central to this perspective are the activities that drive mobility, including economic, leisure, and shopping activities. Accessibility to these activities becomes the key factor in explaining the need for people and goods to travel, bridging the spatial separation of activities resulting from the urban spatial structure determined by land use. The transport infrastructure plays a crucial role in providing the physical connectivity required to bring people to their desired activities.

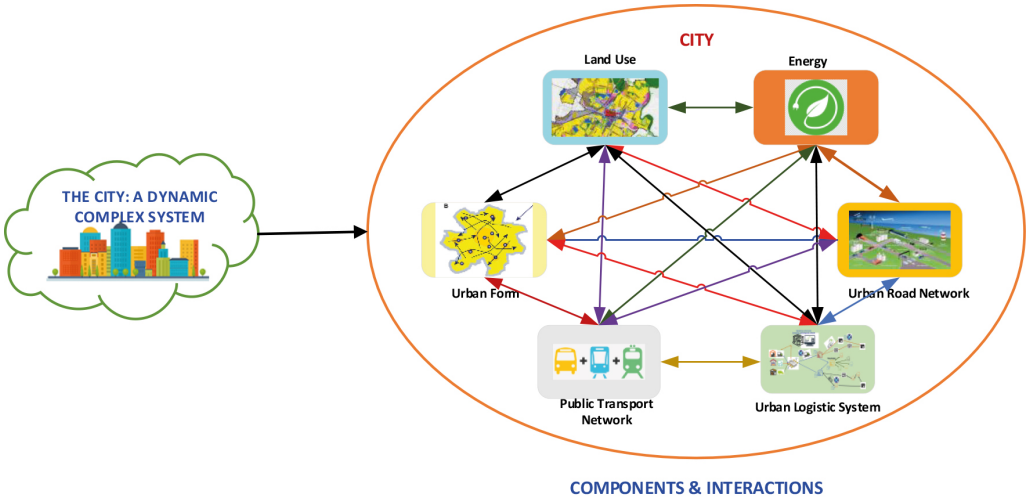

Figure 29. *The complex systems and components of the city.*

However, a holistic perspective cannot overlook the fact that the transport system is reliant on energy, particularly when considering sustainability and the associated energy consumption and emissions from transportation. Shifting towards more sustainable transport technologies, such as replacing fossil fuel-powered vehicles with electric vehicles, necessitates the inclusion of energy grids as part of the system. Figure 29 schematically depicts this approach, presenting the city as a complex system comprising key components: urban form, land use, energy, road network, public transport networks, and the logistics system, which is responsible for ensuring goods reach activity locations but

is often overlooked in conventional approaches. The interactions of these components shape the functioning of the overall system.

This paradigm shift changes the standard modeling approaches and calls for the transition from individual transport system models (e.g., private, public, urban logistics) to a comprehensive model of the city that incorporates urban form, land use, integrated transport networks, energy systems, and charging grids. In recent years, various modeling tools have been developed to support this new approach. One such tool is UrbanSim, an open-source platform designed by Waddell (2011), Waddell (2015), and Waddell et al. (2018). It can integrate with transport planning modeling software like SATURN (Hall, Van Vliet and Willumsen, 1980) and Visum (PTV AG, 2020). Transport modeling has progressed beyond the basic four-step model described in Section 2, which assumes that trips originating from one TAZ and destined for other TAZs are solely determined by the socioeconomic characteristics of the TAZs, implicitly depending on land use. Consequently, changing land use characteristics will also change the number of trips generated in the TAZ. Land use transport integrated (LUTI) models developed by Wegener and Fürst (1999), Acheampong and Silva (2015), and van Wee (2015) explicitly account for these interdependencies. The integration of transport planning software into UrbanSim represents a notable advance in this modeling direction, and early examples can be found in reports from the EU project SIMBAD (Nicolas and Zuccarello, 2011; Dasigi, 2015).

Figure 29, adapted from SIMBAD Project, depicts the conceptual diagram and the flow of information and data between the various modules in this integrated framework, which also includes an urban freight model that addresses the previously overlooked freight traffic flows in conventional models.

A more recent and more powerful software platform for city modeling that integrates ad hoc models for each component is SimMobility (Adnan et al., 2016; Zhu et al., 2018). This software is described on the MIT SimMobility website (https://mfc.mit.edu/simmobility) as follows: "SimMoblity is the simulation platform of the Future Urban Mobility Research Group at the Singapore-MIT Alliance for Research and Technology (SMART) that aims to serve as the nexus of Future Mobility research evaluations. It integrates various mobility-sensitive behavioral models with state-of-the-art scalable simulators to predict the impact of mobility demands on transportation networks, services, and vehicular emissions.

intelligent transportation The platform enables the simulation of the effects of a portfolio of technology, policy, and investment options under alternative future scenarios. Specifically, SimMobility encompasses the modeling of millions of agents, from pedestrians to drivers, from phones and traffic lights to GPS, from cars to buses and trains, from second-by-second to year-by-year simulations, across entire countries". As this presentation highlights, SimMobility offers the additional advantage of being an activity-based approach that fully integrates urban freight transport (Sakaia et al., 2020).

Upon analyzing the dynamics of transportation, it becomes apparent that it is significantly more intricate than the typical simplifications employed in four-step trip-based models. In these models, the trip serves as the fundamental unit of analysis, treating in-

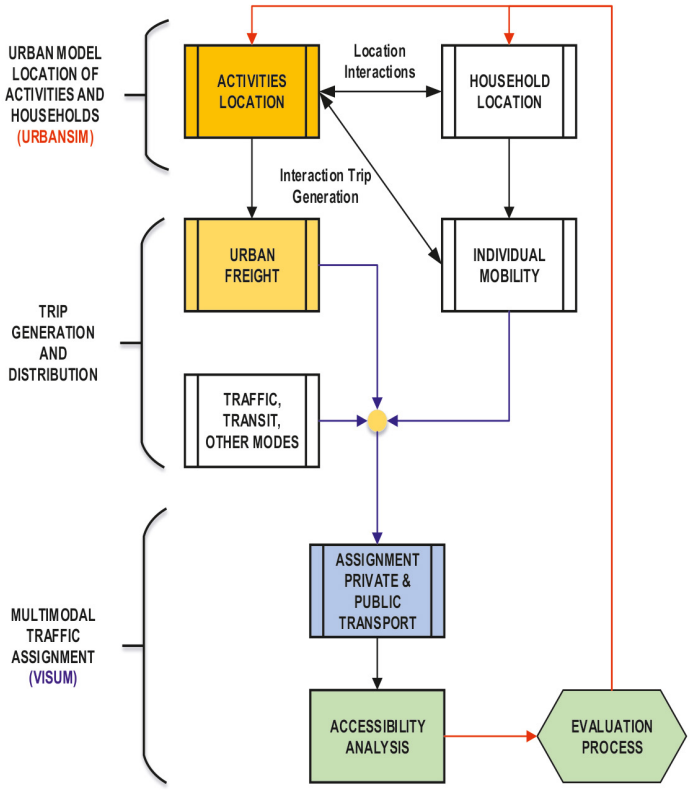

Figure 30. *Conceptual diagram of integrating UrbanSim with the Visum transport model and an urban freight model.*

dividual trips as independent and separate entities. However, when considering the daily schedules of individuals and their activities, we must look at them in terms of sequential chains, like the one illustrated in Figure 31, where a sequence is defined as a series of time points where a person transitions from one discrete state (activity) to another.

In the generic example in Figure 31, the person starts at origin O (home, for example) and travels to activity A_i, perhaps taking their children to school by walking. The duration of this activity is t_i. The person then travels for a duration of τ_{ij} by a transport mode such as a bus from the location of activity A_i to the location of activity A_j, say, to work during duration t_j. He or she then travels by a transport mode that may be the same as or different from the location of activity A_j to the location of activity A_k (say, shopping) over a duration that is τ_{jk} time units. After t_k time units, activity A_k is completed and the individual moves to another destination, D.

From this description, it is clear that the conventional four-step trip-based models lack the necessary structure to represent either a journey's sequential decisions, which now appear as an intermodal chain, or their interrelationships.

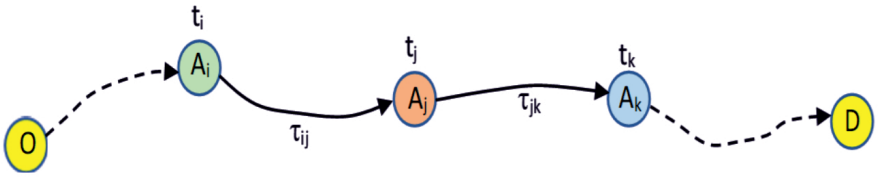

Figure 31. *Examples of activity sequences in a daily schedule.*

The new approach, exemplified by models such as SimMobility and MATSim (Horni, Nagel and Axhausen, 2016), focus on replicating actual traveler decisions by thoroughly understanding the motivations and processes behind them. These models aim to comprehensively represent various interrelated aspects, such as the types of activities individuals engage in, the locations and timing of these activities, and the modes of transportation used to reach them. This entails not only the ability to generate and schedule the activities, which provides insights into the activities people participate in, but also to generate tours and trips with specific destinations and modal choices for reaching them. By considering these factors, the models can identify the routes and modes individuals will utilize, leading to the subsequent network assignment.

Generating the schedule of activities, as depicted in Figure 31, consists of identifying the number and type of activities, their sequential order, the start time and duration of each activity, the modal choices, and the routes taken.

This analysis is conducted through an agent-based simulation, in which individuals are represented as agents whose behaviors are modeled by the decision processes generated through an activity-based approach. Agent-based simulation explicitly incorporates multimodality by simulating the available transportation modes such as cars, buses, and metros, allowing agents to switch between modes according to their schedules.

Additionally, agent-based simulation can effectively address urban freight transport by considering fleets of vehicles and agents in order to schedule their activities as sequences of visits for pickups and deliveries.

Figure 32 depicts the logical diagram of agent-based simulation supported by the activity-based approach. In terms of structural components, it shares similarities with the assignment, mesoscopic, and microscopic models discussed in Section 2. This is because the network supply model, which includes the networks of all available transportation modes, must be built using the same data sources (i.e., GIS and all complementary urban information) that are typically used in transport modeling.

Activity-based models require a huge amount of data, since they must generate information by combining socioeconomic (census tracts) and land-use data with survey data by employing specific sampling techniques like Gibbs sampling to generate synthetic populations, which will in turn be used to generate agents and their activity plans. A seminal work on these applications to agent-based simulation can be found in Farooq et al. (2013), and a more comprehensive overview of available methods can be found

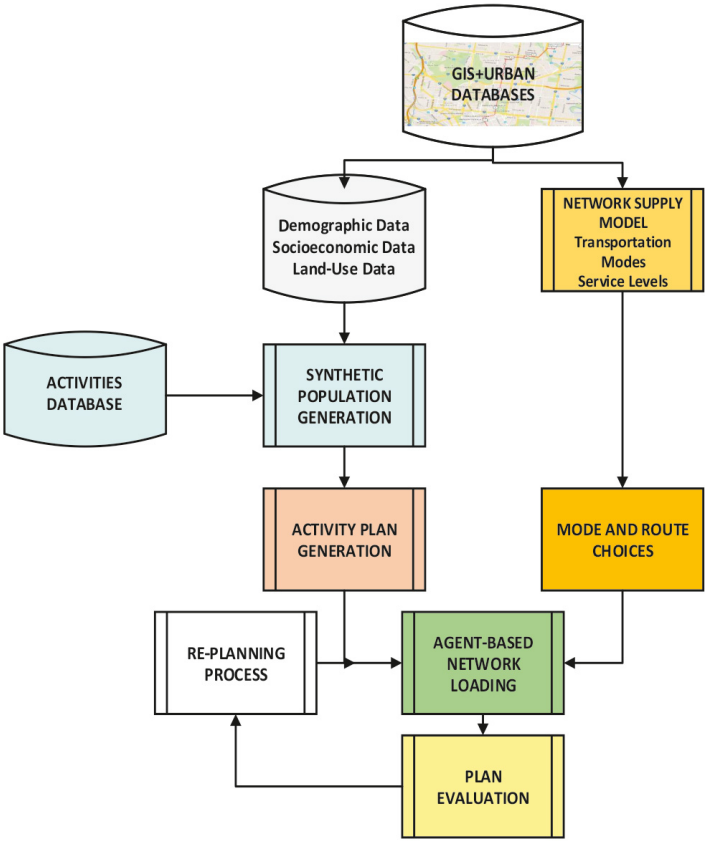

Figure 32. *Activity- and agent-based simulation models.*

in Chapuis, Taillandier and Drogoul (2022). One example of applying MATSim using available data for Barcelona, which was previously discussed in Section 3.2.1 regarding the use of mobile phone data, can be seen in the work of Bassolas et al. (2019). The logic diagram in Figure 30 describes the simulation process as an iterative process, where the performance is evaluated using suitable indicators. Models like SimMobility and MATSim provide sets of indicators, and changes are introduced accordingly, such as adjustments in route or modal choices based on discrete choice models, thereby aiming to achieve some form of equilibrium while emulating individual behaviors.

Funding: This research project has been funded by Spanish R+D Programs, specifically under Grant PID2020-112967GB-C31.

References

Acheampong, R.A. and Silva, E. A. (2015). Land use–transport interaction modeling: A review of the literature and future research directions. *The Journal of Transport and Land Use*, 8(3), 11-38. http://jtlu.org

Ackoff, R.L., Gupta, S.K. and Minas, J.S. (1965). *The Scientific Method. Optimizing Applied Research Decisions*. John Wiley.

Adnan, M., Pereira F.C., Azevedo, C.M.L., Basak, K., Lovric, M. Feliu, S.R., Zhu, Y., Ferreira, J., Zegras, C. and Ben-Akiva, M.E. (2016). *SimMobility: A Multi-Scale Integrated Agent-based Simulation Platform*. 95th 42 Annual Meeting, Transportation Research Board.

Ahmed, K.I. (1999). *Modeling Drivers' Acceleration and Lane Changing Behaviors*. PhD thesis, Massachusetts Institute of Technology.

Akcelik, R. (1991). Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time function. *Australian Road Research*, 21, 49-59.

Alexander, L., Jiang, S., Murga M. and González, M.C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C*, 58, 240-250 http://dx.doi.org/10.1016/j.trc.2015.02.018.

Alexiadis, V., Colyar, J. and Halkias, J. (2007). A Model Endeavor, *Public Roads*, 70(4). Publication Number: FHWA-HRT-07-002.

Anderson, C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. wired.com.

Antoniou, C., Barceló, J., Breen, M., Bullejos, M., Casas, J., Cipriani, E., Ciuffo, B., Djukic, T., Hoogendoorn, S. P., Marzano, V. et al. (2016). Towards a Generic Benchmarking Platform for Origin-Destination Flows Estimation/Updating Algorithms: Design, Demonstration and Validation. *Transp. Res. Part C Emerg. Technol.*, 66, 79-98. https://doi.org/10.1016/j.trc.2015.08.009.

Asakura, Y. and Hato, E. (2009). Tracking Individual Travel Behavior using Mobile Phones: recent technological developments. Chapter 9 in:The Expanding Sphere of Travel Behavior Research. *Selected papers from the 11th International Conference of Travel Behavior Research. Emerald Publishing Limited*. ISBN: 978-1- 84855-936-3.

Asakura, Y. and Hato, E. (2004). Tracking survey for individual travel behaviour using mobile communication instruments. *Transportation Research Part C*, 12, 273-291.

Bando, M., Hasebe, K., Nakanishi, K. and Nakayama, A. (1998). Analysis of optimal velocity model with explicit delay. *Physical Review E*, 58, 5.

Bando, M., Hasebe, K., Nakayama, A., Shibata, A. and Sugiyama, Y. (1995). Dynamic model of traffic congestion and numerical simulation. *Phys. Rev. E*, 51, 1035-1042.

Barceló, J., Ros-Roca, X. and Montero, L. (2022). Data Analytics and Models for Understanding and Predicting Travel Patterns in Urban Scenarios. Chapter 7 in *The Evolution of Travel Time Information Systems*, M. Martínez-Díaz (ed.), Springer Tracts on Transportation and Traffic 19. Doi: 10.1007/978-3-030-89672- 0_6.

Barceló, J. (2019). Future trends in sustainable transportation, Chapter 16 in *Sustainable Transportation and Smart Logistics. Decision Making Models and Solutions*, Editors: J. Faulin, S. Grasman, A. Juan, P. Hirsch, Elsevier, ISBN: 978-0-12-814242-4.

Barceló, J. (2015). Analytics and the Art of Modeling, *International Transactions in Operations Research (ITOR)*, 22, 429-471.

Barceló, J., Gilliéron, F., Linares, M.P., Serch, O. and Montero, L. (2012). Exploring Link Covering and Node Covering Formulations of Detection Layout Problem. *Transportation Research Records: Journal of the Transportation Research Board*, 2308, 17-26.

Barceló, J. (2010). Models, Traffic Models, Simulation and Traffic Simulation, Chapter 1 in J. Barceló, Ed. *Fundamentals of Traffic Simulation*, Springer, ISBN: 978-1-4419-6142-6.

Barceló, J., Montero, L., Marqués, L. and Carmona, C. (2010). Travel time forecasting and dynamic of estimation in freeways based on Bluetooth traffic monitoring, *Transportation Research Records: Journal of the Transportation Research Board*, 2175, 19-27.

Bar-Gera, H. (2002). Origin-based algorithms for the traffic assignment problem. *Transportation Science*, 36, 398- 417.

Bassolas, A., Ramasco, J.J., Herranz, R. and Cantu-Ros, O.G. (2019). Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona. *Transportation Research Part A*, 121 56-74, https://doi.org/10.1016/j.tra.2018.12.024

Behara, K.N.S. (2019). Origin-Destination Matrix Estimation Using Big Traffic Data: A Structural Perspective. *School of Civil Engineering and Built Environment Science and Engineering Faculty Queensland University of Technology*.

Ben-Akiva, M., Koutsopoulos, H.N., Antoniou, C. and Balakrishna, R. (2010). *Traffic Simulation with DynaMIT*. In *Fundamentals of Traffic Simulation*, edited by Barceló, J. Switzerland: Springer, ISBN 978-1-4419-6142-6.

Ben-Akiva, M., Bierlaire, M., Burton, D., Koutsopoulos, H.N. and Mishalani, R. (2001). Network State Estimation and Prediction for Real-Time Traffic Management. *Networks and Spatial Economics*, 1, 293-318.

Ben-Akiva, M. and Bierlaire, M. (1999). Discrete Choice Methods and Their Applications to Short Term Travel Decisions. In Hall, R. W. (ed.). *Handbook of Transportation Science*. http://roso.epfl.ch/mbi/handbook-final.pdf

Ben-Akiva, M. and Lerman, S. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand. *Transportation Studies*. Massachusetts: MIT Press.

Bessa, J., de Magalhães V. and Santos, G.H. (2021). Calibration and Validation of a Volume-Delay Function using Genetic Algorithms. *Journal of Urban and Environmental Engineering*, 15(3), 173-179. DOI:10.4090/juee. 2021.v15n2.173179.

Bevrani, K. and Chung, E. (2011). Car following model improvement for traffic safety metrics reproduction. In *Proceedings of the Australasian Transport Research Forum 2011*, PATREC, Adelaide Hilton Hotel, Adelaid, SA, 1-14.

Blackett, P.M.S. (1948). Operational Research. *The Advancement of Science*, 5(17), 26-38.

Bovy, P., Bekhor, S. and Prato, C. (2008). The factor of revisited path size. *Transp. Res. Board*, 2076, 132-140. http://dx.doi.org/10.3141/2076-15.

Box, G.E.P. and Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces*, John Wiley & Sons.

Box, G.E.P. (1976). *Journal of the American Statistical Association*, 71(356), 791-799.

Bunge, M. (1960). *La ciencia, su método y su filosofía*. Buenos Aires : Eudeba. (In French : *La science, sa méthode et sa philosophie*. Paris: Vigdor, 2001. [ISBN 2-910243-90-7]).

Bureau of Public Roads (1964). *Traffic Assignment Manual*. Urban Planning Division, US Department of Commerce, Washington, DC.

Butterworth, S.(1930). On the Theory of Filter Amplifiers. *Experimental Wireless and the Wireless Engineer.*, 7, 536- 541. https://www.changpuak.ch/electronics/downloads/On_the_Theory_of_Filter_Amplifiers.pdf

Cascetta, E. (2001). *Transportation System Engineering: Theory and Methods*. Kluwer, ISBN: 0-7293-6972-8.

Cascetta, E., Nuzzolo, A., Russo, F. and Vitetta, A. (1996). A modified logit route choice model overcoming path overlapping problems. In: *Proceedings of the 13th International Symposium on the Theory of Road Traffic Flow*.

Castillo, E., Conejo, A.J., Menéndez, J.M. and Jiménez, P. (2008). The Observability Problem in Traffic Network Models. *Computer-Aided Civil and Infrastructure Engineering*, 23, 208-222.

Chandler, R.E., Herman, R. and Montroll, E.W. (1958). Traffic Dynamics: Studies in Car Following. *Opns. Res.*, 6, 165-184.

Chapuis, K., Taillandier, P. and Drogoul, A.(2022). Generation of Synthetic Populations in Social Simulations: A Review of Methods and Practices. *Journal of Artificial Societies and Social Simulation*, 25(2), 6. Doi: 10.18564/jasss.4762 Url: http://jasss.soc.surrey.ac.uk/25/2/6.html

Chiu, Y.C., Bottom, J., Mahut, M., Paz, A., Balakrishna, R., Waller, T. and Hicks, J. (2011). Dynamic traffic assignment: a primer. *Transportation Research E-Circular*, (E-C153).

Cluet, J. (2021). Analysis of GPS tracking data for traffic modelling parameters estimation. Bachelor Thesis report. *Computer Science Degree (GEI)*, Facultat d'InformÃ tica de Barcelona, Universitat PolitÃ¨cnica de Catalunya.

Codina, E., Ibáñez, G. and Barceló, J. (2015). Applying Projection-based Methods to the Asymmetric Traffic Assignment Problem. *Computer-Aided Civil and Infrastructure Engineering*, 20(2), 103-119.

Coifman, B. (2017). *Data Sets*. https://www.ece.osu.edu/?coifman/documents.

Coifman, B. and Li, L. (2017). A critical evaluation of the Next Generation Simulation (NGSIM) vehicle trajectory dataset. *Transportation Research Part B*, 105, 362-377.

Çolak, S., Lima, A. and González, M.C. (2016). Understanding congested travel in urban areas. *Nature Communications*. DOI: 10.1038/ncomms10793

Daganzo, C.F. (1997). *Fundamentals of transportation and traffic operations*. Pergamon Oxford.

Daganzo, C.F. (1994). The cell-transmission model: a simple dynamic representation of highway traffic. *Transportation Research Part B: Methodological*, 28(4), 269–287.

Daganzo, C.F. (1995a). The cell transmission model part II: network traffic. *Transportation Research Part B: Methodological*, 29, 79-93.

Daganzo, C.F. (1995b). A finite difference approximation of the kinematic wave model of traffic flow. *Transportation Research Part B: Methodological*, 29(4), 261-276.

Dai, Y., Yang, Y., Zhong, H., Zuo, H. and Zhang, Q. (2022). Stability and Safety of Cooperative Adaptive Cruise Control Vehicular Platoon under Diverse Information Flow Topologies. *Hindawi, Wireless Communications and Mobile Computing*, Volume 2022, Article ID 4534692, https://doi.org/10.1155/2022/4534692

Dailey, D.H., Harn, P. and Lin, P-J. (1996). Final Research Report, Research Project T9903, Task 9, ATIS/ATMS Regional IVHS Demonstration, prepared for *Washington State Transportation Commission*, Department of Transportation.

Dasigi, S. (2015). *An integrated approach linking Land Use and socioeconomic characteristics for improving travel demand forecasting*. University of Cincinnati.

de Montjoye, Y-A., Smoreda, Z., Trinquart, R., Ziemlicki, C. and Blondel, V. D. (2014). *D4D-Senegal: The Second Mobile Phone Data for Development Challenge*, http://arxiv.org/pdf/1407.4885v2.

Dial, R. (2006). A path-based user equilibrium traffic assignment algorithm that obviates paths storage and enumeration. *Transportation Research Part B*, 40, 917-936.

Djukic, T. (2014). Dynamic OD Demand Estimation and Prediction for Dynamic Traffic Management, TU Delft, 2014. https://doi.org/10.4233/uuid:ab12d7a7-e77b-424d-b478-d58657f94dd1

Edie, L. (1963). Discussion of traffic stream measurements and definitions, in: Almond, J. (Ed.). *Proceedings of the 2nd International Symposium on the Theory of Traffic Flow*, 139-154.

Farooq, B., Bierlaire, M., Hurtubia, R. and Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58, 243-263.

Florian, M. and Constantin, I. (2009). A new look at the projected gradient method for equilibrium assignment. *Transportation Research record Journal of the Transportation research Board*. Paer 2090, pp. 10-16. DOI:10.3141/2090-02.

Florian, M., Mahut, M. and Tremblay, N. (2001). A Hybrid Optimization-Mesoscopic Simulation Dynamic Traffic Assignment Model. *Proceedings of the 2001 IEEE Intelligent Transport Systems Conference*, 2001, Oakland, 118-123.

Florian, M., Mahut, M. and Tremblay, N. (2002). Application of a Simulation-Based Dynamic Traffic Assignment Model. *International Symposium on Transport Simulation*, 2002, Yokohama. (also in: *Simulation Approaches in Transportation Analysis*, 2005, edited by R. Kitamura and M. Kuwahara, US: Kluwer.)

Florian, M. and Hearn, D. (1995). Network Equilibrium Models and Algorithms, Chapter 6 in: M.O. Ball et al., Eds., *Handbooks in OR and MS*, Vol.8, Elsevier Science B.V.

Florian, M., Guelat, J. and Spiess, H. (1987). An Efficient Implementation of the PARTAN Variant of the Linear approximation Method for the Network Equilibrium Problem. *Networks*, 17, 319-339.

Florian, M. and Nguyen, S. (1976). An application and validation of Equilibrium trip assignment models. *Transportation Science*, 10(4), 374-390.

Frank, M. and Wolfe, P. (1956). An Algorithm for Quadratic Programming. *Naval Researchs Logistic Quarterly*, 3, 95-110.

Fréchet, M.M. (1906). Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo*, 22(1), 1-72.

Frederix, R., Viti, F. and TampÃ¨re, C.M.J. (2013). Dynamic Origin-Destination Estimation in Congested Networks: Theoretical Findings and Implications in Practice. *Transp. A Transp. Sci.*, 9(6), 494-513. https://doi.org/10.1080/18128602.2011.619587.

Friesz, T.L., Bernstein, D., Smith, T.E., Tobin, R.L. and Wie, B.W. (1993). A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem. *Operations Research*, 41(1), 179-191.

Fritzsche, H.T. (1994). A model for traffic simulation. *Transp. Engin. Contr.*, 5, 317-321.

García-Albertos, P., Picornell, M., Salas-Olmedo, M. and Gutiérrez, J. (2018). Exploring the potential of mobile phone records and online route planners for dynamic accessibility analysis. *Transportation Research Part A*.

Gazis, D.C., Herman, R.L. and Rothery, R. W. (1961). Nonlinear Follow-the-Leader Models of Traffic Flow. *Operations Research*, 9(4), 545-567.

Gazis, D.C., Herman, R. and Potts, R.B. (1959). Car-Following Theory of Steady- State Traffic Flow. *Opns. Res.*, 7, 499-505.

Gentile, G. (2010). *The General Link Transmission Model for Dynamic Network Loading and a comparison with the DUE algorithm*. In *New Developments In Transport Planning: Advances In Dynamic Traffic Assignment*, edited by Immers, L.G.H., Tampere C.M.J. and F. Viti. MA, USA: Transport Economics, Management and Policy Series, Edward Elgar Publishing.

Gentile, G. (2015). Using the General Link Transmission Model in a Dynamic Traffic Assignment to simulate congestion on urban networks. *Transportation Research Procedia,* 5, 66-81.

Gerlough, D.L. and Huber, M.J. (1975). Traffic Flow Theory: A Monograph. *TRB Special Report*, 165.

González, M.C., Hidalgo, C.A. and Barabási, A.L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196), 779-782.

Guerrero-Ibáñez, J., Zeadally, S. and Contreras-Castillo, J. (2018). ISensor Technologies for Intelligent Transportation Systems. *Sensors 2018*, 18, 1212. doi:10.3390/s18041212

Gundlegård, D., Rydergren, C., Breyer, N. and Rajna, B. (2016). Travel demand estimation and network assignment based on cellular network data. *Computer Communications*, 95, 29-42. http://dx.doi.org/10.1016/j.comcom.2016.04.015

Gundlegård, D., Rydergren, C., Barcelo, J., Dokoohaki, N., Görnerup, O. and Hess, A. *Travel Demand Analysis with Differentially Private Releases*. Submitted for D4D Challenge Senegal 2014 (Netmob 2015, MIT, Boston).

Gipps, P.G. (1981). A behavioral car-following model for computer simulation. *Transp. Res. B*, 15, 105-111.

Hall, M. D., Van Vliet, D. and Willumsen, L.(1980). SATURN—a simulation-assignment model for the evaluation of traffic management schemes. *Traffic Engineering and Control*, 21(4), 168-176.

Hariharan, R. and Toyama, K. (2004). Project Lachesis: Parsing and Modeling Location Histories. In M.J. Egenhofer, C. Freksa, and H.J. Miller (Eds.): *GIScience 2004*, LNCS 3234, 106-124, 2004. Springer Verlag.

Hato, E (2010). Development of behavioral context addressable loggers in the shell for travel-activity analysis. *Transportation Research Part C*, 1855-67.

Hearn, D.W., Lawphonpanich, S. and Ventura, J.A. (1987). Restricted Simplicial Decomposition: Computation and Extensions. *Mathematical Programming Study*, 31, 99-118.

Heavens, N.G., Ward, D.S. and Natalie, M.M. (2013).Studying and Projecting Climate Change with Earth System Models. *Nature Education Knowledge*, 4(5), 4.

Herman, R., Montroll, E.W. and Potts, R.B. (1959). Traffic Dynamics: Analysis of Stability in Car Following. *Opns. Res.*, 7, 86-106.

Horni, A., Nagel, K. and Axhausen, K.W. (eds.) (2016). *The Multi-Agent Transport Simulation MATSim. London: Ubiquity Press*. DOI: http://dx.doi.org/10.5334/baw.

Iqbal Md. S., Choudhury, C.F., Wang, P. and Gonzalez, M.C. (2014). Development of Origin-Destination Matrices Using Mobile Phone Call Data. *Transportation Research Part C*, 40, 63-74.

ITF (2015). Big Data and Transport: Understanding and Assessing Options. *International Transport Forum*. OECD.

Itoh, S. and Hato, E. (2013). Combined estimation of activity generation models incorporating unobserved small trips using probe person data. *Journal of the Eastern Asia Society for Transportation Studies*, 10, 525-537.

Janmyr, J. and Wadell, D. (2018). *Analysis of Vehicle Route Choice During Incidents* (MSc thesis). University of Linköping, Department of Science and Technology, LiU-ITN-TEK-A–18/020-SE.

Jayakrisham, R., Mahmassani, H.S. and Yu, T.Y. (1994). An Evaluation Tool for Advanced Traffic Information and Management Systems in Urban Networks. *Transportation Research Part C: Emerging Technologies*, 2C(3), 129-147.

Jiang, S., Yanga, Y., Gupta, S., Veneziano, D., Athavale, S. and González, M. C. (2016). Supporting Information. *PNAS 2016*, vol. XXX, no. XX , 1-11.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problem. *Journal of Basic Engineering*, 82(1), 35-45.

Kesting, A., Treiber, M. and Helbing, D. (2010). Enhanced Intelligent Diver Model to Access the Impact of Driving Strategies on Traffic capacity Simulations. *Philosophical Transactions of the Royal Society A*, 368, 4585-4605.

Kim, I-S. and McLean, W. (2013). Computing the Hausdorff distance between two sets of parametric curves. *Communications of the Korean Mathematical Society*. Oct 31; 28(4), 833-50. https://doi.org/10.4134/CKMS.2013.28.4.833

Kotsialos, A. and Papageorgiou, M. (2001). The Importance of Traffic Flow Modeling for Motorway Traffic Control. *Networks and Spatial Economics*, 1, 179-203.

Krauss, S., Wagner, P. and Gawron, C. (1997). Metastable States in a Microscopic Model of Traffic Flow. *Physical Review E*, 55(5), 55-97.

Krishnakumari, P., Van Lint, H., Djukic, T. and Cats, O. (2019). A data driven method for OD matrix estimation. *Transportation Research C*, doi: 10.1016/j.trc.2019.05.014.

Kubicka, M., Cela, A., Mounier, H. and. Niculescu, S.I. (2018). *Comparative Study and Application-Oriented Classification of Vehicular Map-Matching Methods*. doi: 10.1109/MITS.2018.2806630

Lawphongpanich, S. and Hearn, D.W. (1984). Simplicial Decomposition of the Asymmetric Traffic Assignment Problem. *Transportation Research 18B*, 123-133.

LeBlanc, L.J., Morlok, E.K. and Pierskalla, W.P. (1975). An Efficient Approach for Solving the Road Network Equilibrium Traffic Assignment Problem. *Transportation Research*, 5, 309-318.

Levinson, D.M. and Kumar, A. (1994). The rational locator: why travel times have remained stable. *Journal of the American Planning Association*, 60(3), 319-332.

Lighthill, M. and Whitham, G. (1955). On Kinematic Waves II: A Traffic Flow Theory on Long Crowded Roads. *Proceedings of the Royal Society of London Series A*, 229, 317-345.

Lopez, C., Krishnakumari, P., Leclercq, L., Chiabaut, N. and Van Lint, H. (2017a). Spatiotemporal Partitioning of Transportation Network Using Travel Time Data. *Transportation Research Record: Journal of the TRB*, 2623, 98-107.

Lopez, C., Leclercq, L., Krishnakumari, P., Chiabaut, N. and Van Lint, H. (2017b). Revealing the day-to-day regularity of urban congestion patterns with 3D speed maps. *Scientific Reports*, 2(7), 14029.

Lu, X-Y. and Skabardonis, A. (2007). Freeway Traffic Shockwave Analysis: Exploring the NGSIM Trajectory Data. *TRB 86th Annual Meeting Compendium of Papers CD-ROM*, Paper #07-3016.

Mahmassani, H. (2001). Dynamic Network Traffic Assignment and Simulation Methodology for Advanced System Management Applications. *Network and Spatial Economics*, 1, 267-292.

Mahmassani, H.S., Hu, T.Y., Peeta, S. and Ziliaskopoulos, A. (1994). Development and Testing of Dynamic traffic assignment and Simulation procedures for ATIS/ATMS

Applications. Technical Report DTFH61-90-R00074-FG, Center for Transportation research, The University of Texas at Austin.

Mahut, M. and Florian, M. (2010). *Traffic Simulation with Dynameq.* In *Fundamentals of Traffic Simulation*, edited by Barceló, J. Switzerland: Springer, ISBN 978-1-4419-6142-6.

Mahut, M. (1999). *Behavioural Car Following Models. Report CRT-99-31*. Centre for Research on Transportation. University of Montreal. Montreal, Canada.

Mahut, M. (2001). *Discrete flow model for dynamic network loading*. Ph.D. Thesis, Département d'Informatique et de Recherhe Opérationelle, Université de Montréal. Published by the Center for Research on Transportation (CRT), University of Montreal.

McNally, M.G. (2000). *The Four Step Model*. UCI-ITS-AS-WP-00-5. Institute of Transportation Studies and Department of Civil & Environmental Engineering University of California, Irvine. http://www.its.uci.edu

Michalopoulos, P.G., Yi, P. and Lyrintzis, A.S. (1993). Continuum modelling of traffic dynamics for congested freeways. *Transportation Research Part B: Methodological*, 27(4), 315-332.

Michalopoulos, P.G. (1984). Dynamic Freeway Simulation Program for Personal Computers. In *Transportation Research Records 971, National Research Council*, Washington, D.C., 68-79.

Michalopoulos, P.G., Beskos, D.E. and LinJ, K. (1984). Analysis of Interrupted Traffic Flow by Finite Difference Methods. *Transportation Research*, 18B, 409-421.

Millard-Ball, A., Hampshire, R.C. and Weinberger, R.R. (2019). Map-matching poor-quality GPS data in urban environments: the pgMapMatch package. *Transp. Plan. Technol.*, 42(6), 539-553. doi: 10.1080/03081060.2019.1622249.

Minsky, M. (1965). Matter, Mind and Models. In *Proceedings of IFIP Congress*, edited by I. F. of Information Processing Congress, 45-49.

Mitchell, R.B. and Rapkin, C. (1954). *Urban Traffic: A Function of Land Use*, Columbia University Press, New York, NY.

Mitra, A., Attanasi, A., Meschini, L. and Gentile, G. (2020). Methodology for O-D Matrix Estimation Using the Revealed Paths of Floating Car Data on Large-scale Networks. *IET Intell. Transp. Syst.*, 14(12), 1704-1711.

Montanino, M. and Punzo, V. (2015). Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns. *Transportation Research Part B*, 80, 82-106.

Montero, L., Ros-Roca, X., Herranz, R. and Barceló, J. (2019). Fusing mobile phone data with other data sources to generate input OD matrices for transport models. *Transportation Research Procedia* (www.sciencedirect.com), 37, 417-424.

MULTITUDE (2014). *Traffic Simulation: Case for Guidelines*. COST Action TU0903 MULTITUDE, European Commission Joint Research Centre Institute for Energy and Transport, http://iet.jrc.ec.europa.eu/

Newell, G.F. (1961). Nonlinear Effects in the Dynamic of Car-Following. *Operations Research*, 9(2).

Nicolas, J.P. and Zuccarello, P. (2011). Land-Use and transport Interaction Modeling: The SIMBAD Project and the Articulation between Visum and URBANSIM. *PTV Vision International Users Group Meeting*, New York, September 2011

OECD/ITF (2015). *Big data and Transport: Understanding and Assessing Options*. International Transport Forum. https://www.oecd-ilibrary.org/transport/big-data-and-%09transport_5jlwvzdb6r47-enhttps://www.oecd-ilibrary.org/transport/big-data-and-transport_5jlwvzdb6r47-en

Ortúzar, J.D. and Willumsen, L.G. (2011). *Modelling Transport*, John Wiley.

Osorio, C. (2019). Dynamic Origin-Destination Matrix Calibration for Large-Scale Network Simulators. *Transp. Res. Part C Emerg. Technol.*, 98 (April 2018), 186-206. https://doi.org/10.1016/j.trc.2018.09.023.

Papageorgiou, M. (1998). Some remarks on macroscopic traffic flow modelling. *Transportation Research Part A: Policy and Practice*, 32(5), 323-329.

Papageorgiou, M., Blosseville, J. and Hadj-Salem, H. (1990). Modelling and Real-Time Control of Traffic Flow on the Southern Part of Boulevard Peripherique in Paris. *Part I: Modelling, Transportation Research A*, 24, 345-359.

Papageorgiou, M., Blosseville, J.M. and Hadj-Salem, H. (1989). Macroscopic modelling of traffic flow on the Boulevard Périphérique in Paris. *Transportation Research Part B: Methodological*, 23(1), 29-47.

Patriksson, M. (1994). *The Traffic Assignment Problem*, VSP. (the pdf of this book is available on Patriksson's web page at http://www.cs.chalmers.se/~mipat/traffic.html

Payne, H. (1979). FREFLO: A macroscopic simulation model of freeway traffic. *Transportation Research Record*, 772, 68-75.

Payne, H.J. (1971). Models of freeway traffic and control. *Mathematical Models of Public Systems*, 1(1), 51-61.

Petrik, O., Moura, F. and Abreu e Silva, J. (2014). The Influence of the Volume-Delay Function on Uncertainty Assessment for a Four-Step Model. *Advances in Intelligent Systems and Computing*, 262, 293-306. DOI: 10.1007/978-3- 319-04630-3 22.

Pipes, L.A. (1953). An Operational Analysis of Traffic Dynamics. *Journal of Applied Physics*, 24(3).

Popper, K.R. (1972). *Conjectures and Refutations. The Growth of Scientific Knowledge*. Routledge & Kegan, London.

PTV AG (2020). PTV Visum 2020 - User's Manual. PTV Group, Karlsruhe, Germany.

Punzo, V., Borzacchiello, M.T. and Ciuffo, B. (2011). On the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data. *Transportation Research Part C*, 19, 1243-1262.

Ran, B. and Boyce, D. (1996). *Modeling Dynamic Transportation Networks*, Springer-Verlag.

Ratti, C., Frenchman, D., Pulselli, R.M. and Williams, S. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and planning B: Planning and design*, 33(5), 727-748.

Richards, P. (1956). Shock Waves on the Highway. *Operations Research*, 4, 42-51.

Rosen, B. (1960). The gradient projection method for nonlinear programming I. Linear Constraints. *J. Soc. Indust. Appl. Math.*, 9, 181-217.

Ros-Roca, X., Montero, L., Barceló, J., Nökel, K. and Gentile, G. (2022). A practical approach to assignment-free Dynamic Origin–Destination Matrix Estimation problem. *Transportation Research Part C*, 134, https://doi.org/10.1016/j.trc.2021.103477

Ros-Roca, X., Montero, L. and Barceló, J. (2020). Investigating the quality of Spiess-like and SPSA approaches for dynamic od matrix estimation. *Transportmetrica*. https://doi.org/10.1080/23249935.2020.1722282.

Rothery, R.W. (2001). *Car-Following Models. Chapter 4 in TRB (2001)*. Revised Monograph on Traffic Flow Theory.

Rouphail, N.M. and Sacks, J. (2003). *Thoughts on Traffic Models Calibration and Validation*, paper presented at the Workshop on Traffic Modeling, Sitges, Spain.

Sakaia, T., Alho, A.R., Bhavathrathan, B.K., Dalla Chiara, G., Gopalakrishnan, R., Jing, P., Hyodo, T., Cheah, L. and Ben-Akiva, M. (2020). SimMobility Freight: An agent-based urban freight simulator for evaluating logistics solutions. *Transportation Research Part E*, 141. https://doi.org/10.1016/j.tre.2020.102017

Savitzky, A. and Golay, M.J.E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8), 1627-39.

Schafer, A. (2000). Regularities in travel demand: an international perspectiv. *Journal of transportation and statistics*, 3(3), 1-31.

Sheffi, Y. (1985). *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice Hall. (Available from Prof. Sheffi's web site: http://mit.edu/sheffi/www/urbanTransportation)

Smith, M.J. (1993). A new dynamic traffic model and the existence and calculation of dynamic user equilibria on congested capacity-constrained road networks. *Transportation Research Part B*, 27, 49-63.

Smith, M.J. (1979). Existence, Uniqueness and Stability of Traffic Equilibria. *Transportation Research B*, 1B, 295-304.

Spall, J.C. (1992). Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation. *IEEE Trans. Automat. Contr.*, 37(3), 332-341. https://doi.org/10.1109/9.119632.

Spiess, H. (1990). Conical volume delay functions. *Transportation Science*, 24(2), 153-158.

Stock, W.A., Blankenhorn, R.C. and May, D.C. (1973). *The FREQ3 Freeway Model*, ITTE Report 73-2, University of California Berkeley

Talebpour, A. and Mahmassani, H.S. (2016). Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transportation Research Part C*, 1, 143-163.

Talebpour, A., Mahmassani, H.S. and Bustamante, F.E. (2016). Modeling Driver Behavior in a Connected Environment Integrated Microscopic Simulation of Traffic and Mobile Wireless Telecommunication Systems. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2560, Transportation Research Board, Washington, D.C., 2016, 75-86. DOI: 10.3141/2560-09

Toledo, T. and Kolechkina, T. (2013). Estimation of Dynamic Origin-Destination Matrices Using Linear Assignment Matrix Approximations. *IEEE Trans. Intell. Transp. Syst.*, 14(2), 618-626. https://doi.org/10.1109/TITS.2012.2226211.

Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A. and González M.C. (2015). The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C*, 58, 162-177.

TRB (2001). Revised Monograph on *Traffic Flow Theory*. https://www.fhwa.dot.gov/publications/research/operations/tft/

Treiber, M. and Kesting, A. (2013). *Traffic Flow Dynamics. Data, Models and Simulation*. Springer ISBN 978-3-642- 32459-8.

van Wee, B. (2015). Viewpoint: Toward a new generation of land use transport interaction models. *The Journal of Transport and Land Use*, 8(3), 1-10. http://jtlu.org

Viterbi, A.J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2), 260-269.

Waddell, P. (2015). Integrated transport and land use modeling for sustainable cities. Michel Bierlaire, Andrè De Palma, Ricardo Hurtubia, Paul Waddell. EPFL Press.

Waddell, P., Boeing, G., Gardner, M. and Porter, E. (2018). An Integrated Pipeline Architecture for Modeling Urban Land Use. *Travel Demand, and Traffic Assignment*. arXiv:1802.09335v1 [cs.CY].

Waddell, P. (2011). Integrated Land Use and Transportation Planning and Modelling: Addressing Challenges in Research and Practice. *Transport Reviews: A Transnational Transdisciplinary Journal*, 31(2), 209-229, DOI: 10.1080/01441647.2010.525671

Wagner, P. (2016). Traffic control and traffic management in a transportation system with autonomous vehicles. Ch. 15 in M. Maurer et al. (eds). *Autonomous Driving*. DOI 10.10007/978-3-662-48847-8-15.

Ward, J.A. (2009). *Heterogeneity, Lane-Changing and Instability in Traffic: A Mathematical Approach*. Department of Engineering Mathematics. University of Bristol, Bristol, United Kingdom.

Ward, J.A. and Wilson, R.E. (2011). Criteria for convective versus absolute string instability in car-following models. *Proc. R. Soc. A*, 467, 2185-2208. doi:10.1098/rspa.2010.0437

Wardrop, J.G. (1952). Some Theoretical Aspects of Road Traffic Research. *Proc. Inst. Civil Engineers, Part II*, 325-378.

Wegener, M. and Fürst, F. (1999). Land-Use Transport Interaction: State of the Art. Deliverable 2a of the project TRANSLAND (Integration of Transport and Land Use Planning) of the 4th RTD Framework Programme of the European Commission. Institut für Raumplanung. Fakultä Raumplanung, Universität Dortmund.

Wiedemann, R. (1974). Simulation des Verkehrsflusses. Schriftenreihe des Instituts für Verkehrwessen, Heft 8, Universität (TH) Karlsruhe.

Wilson, E. (2011). Mathematical Analysis of CF Models, NEARCTIS Course (Delft 2011).

Wu, J.H. (1991). *A Study of Monotone Variational Inequalities and their Application to Nertwork Equilibrium Problems*, Ph. D. Thesis, Centre de Recherche sur les Transports, Université de Montréal, Publication #801.

Wu, J.H., Chen, Y. and Florian, M. (1998a). The Continuous Dynamic Network Loading Problem: A Mathematical Formulation and Solution Method. *Trans. Res.-B*, 32(3), 173-187.

Wu, J.H., Florian, M., Xu, Y.W. and Rubio-Ardanaz, J.M. (1998b). A projection algorithm for the dynamic network equilibrium problem. *Traffic and Transportation Studies, Proceedings of the ICTTS'98*, 379-390, Ed. By Zhaoxia Yang, Kelvin C.P. Wang and Baohua Mao, ASCE.

Wu, C., Kreidiehy, A., Parvate, K., Vinitskyz, E. and Bayen, A. (2017). Flow: Architecture and Benchmarking for Reinforcement Learning in Traffic Control. arXiv:1710.05465v1 [cs.AI]

Yeo, Y. and Skabardonis, A. (2007). TO 9: Oversaturated Freeway Flow Algorithm, Final Report, University of California Berkeley, NGSIM Program.

Zhang, Y., Qin, X., Dong, S. and Ran, B. (2010). Daily O-D Matrix Estimation Using Cellular Probe Data, Paper 10- 2472, presented at the 89th TRB Annual Meeting, included in the Compendium of Papers.

Zhu, Y., Ferreira, J., Diao, M. and Zegras, P.C. (2018). An integrated microsimulation approach to land-use and mobility modeling. *The Journal of Transport and Land Use*, 11(1), 633-659. http://jtlu.org

SORT 47 (1) January-June 2023, 81-128 DOI: 10.57645/20.8080.02.2

Data science, analytics and artificial intelligence in e-health: trends, applications and challenges

Juliana Castaneda[1], Laura Calvet[2], Sergio Benito[3], Abtin Tondar[4] and Angel A. Juan[5]

Abstract

More than ever, healthcare systems can use data, predictive models, and intelligent algorithms to optimize their operations and the service they provide. This paper reviews the existing literature regarding the use of data science/analytics methods and artificial intelligence algorithms in healthcare. The paper also discusses how healthcare organizations can benefit from these tools to efficiently deal with a myriad of new possibilities and strategies. Examples of real applications are discussed to illustrate the potential of these methods. Finally, the paper highlights the main challenges regarding the use of these methods in healthcare, as well as some open research lines.

MSC: *68T01, 62-07.*

Keywords: *e-health, data science, analytics, artificial intelligence, machine learning.*

1. Introduction

In recent years, there has been a growing trend to digitize much of the data that used to be stored in hard copies. The healthcare industry, which has always been character-

[1] Dept. of Computer Science, Multimedia and Telecommunication, Universitat Oberta de Catalunya, 08018 Barcelona, Spain.

[2] Dept. of Computer Science, Multimedia and Telecommunication, Universitat Oberta de Catalunya, 08018 Barcelona, Spain.

[3] Dept. of Computer Science, Multimedia and Telecommunication, Universitat Oberta de Catalunya, 08018 Barcelona, Spain.

[4] Dept. of Computer Science, Multimedia and Telecommunication, Universitat Oberta de Catalunya, 08018 Barcelona, Spain.

[5] Dept. of Statistics and Operations Research, Universitat Politècnica de València, 03801 Alcoy, Spain.

Received: November 2021
Accepted: July 2022

ized by the generation of large amounts of data, has already begun this digital transformation (Raghupathi and Raghupathi, 2014). The term electronic health, or 'e-health', appeared in the 1990s by the influence of the Internet. The prefix 'e-' became popular to accompany different terms, such as e-mail or e-commerce, referring to various developments in information and communication technology (ICT). The term e-health has been defined by Eysenbach (2001) as: "an emerging field in the intersection of medical informatics, public health and business, referring to health services and information delivered or enhanced through the Internet and related technologies. In a broader sense, the term characterizes not only a technical development, but also a state-of-mind, a way of thinking, an attitude, and a commitment for networked, global thinking, to improve health care locally, regionally, and worldwide by using information and communication technology". This information could include several health-related concepts, as well as various stakeholders, roles, locations, and benefits (Oh et al., 2005). In this sense, ICTs are supporting tools for health-related activities. Within the main domains of e-health, we can find: telemedicine, clinical information systems, different types of medical networks, disease registries for different purposes (education, public health, patient/disease behavior, and healthcare management), mobile health, personalized health, and big data (Cowie et al., 2016). All in all, this allows us to consider e-health as a big data source, and a potential field for applying data science/analytics techniques, including predictive models, optimization algorithms, modeling and simulation, or any other technique that involves data processing for a defined purpose. In addition, new models, such as learning healthcare systems, have been recently developed to facilitate using medical data for improving healthcare (Enticott, Johnson and Teede, 2021).

Data science/analytics emerged as a hybridization of several disciplines, such as statistics, operations research/management science, data mining, computer science, data bases, machine learning, mathematics, and distributed systems. The combination of all the existing methodologies in this field makes the large amounts of data available valuable for individuals, organizations, and society (Van Der Aalst, 2016). Concepts such as artificial intelligence (AI), machine learning (ML), statistical learning, or data science are clearly interconnected, and they share many methods and techniques. For instance, it is possible to find predictive, regression, classification, and clustering models in all the previous concepts. Still, they are not exactly the same concept. Hence, AI is a wide area, being its main goal the development of machines capable of emulating human intelligence. With that purpose, it uses computer science algorithms (e.g., optimization and searching algorithms), statistical methods, ML models, computer vision techniques, etc. ML is usually seen as a subset of AI, and it focuses on a series of supervised methods (classification, regression, predictive models, etc.), unsupervised methods (clustering, dimensionality reduction, etc.), and reinforcement learning methods. Many of these methods are also employed in statistical learning. However, statistical learning is more focused on the statistical fundamentals of these methods, while ML is more oriented towards their computational and programming aspects. Data science, on the other hand also refers to many of the ML and AI methods as well as to the use of databases with

large amounts of data, data gathering and pre-processing, and other analytical methods and algorithms (including, for instance, simulation models, time series analysis, etc.).

Managing big data in healthcare is a complex task because of its volume and the diversity of data types, and the speed at which they must be processed. There is an opportunity for data analysts to discover associations and understand patterns and trends within the data. Big data analytics can improve care, save lives, and reduce costs (Raghupathi and Raghupathi, 2014). Potential applications of data science/analytics methods and AI algorithms to e-health are almost unlimited: from the enhancement of interoperability in e-health systems (Gupta and Gupta, 2019; Razzaque and Hamdan, 2020) to the use of Internet of things (IoT) and algorithms for generating smart healthcare networks (Syed et al., 2019). This includes the use of healthcare data to provide smart medical services to citizens (Haldorai, Ramu and Murugan, 2019) or high-risk pregnancy home healthcare (Moreira et al., 2018). Typical applications of analytics/operations research/management science methods in healthcare –including the pharmaceutical industry as well– can be found in Beheshtifar and Alimoahmmadi (2015), Rais and Viana (2011), Saranga and Phani (2009), or Ahsan and Bartlema (2004).

Based on the Google Scholar and Elsevier Scopus databases, Figure 1 shows the time evolution of the number of scientific articles that contain all the following terms: "artificial intelligence", "e-health", and "data science". Notice the fast growth in the number of papers that combine all the aforementioned terms, which shows a clear tendency in the literature to consider the combined use of data analysis methods and AI algorithms in the e-health sector. Accordingly, one of the main goals of this work is to analyze the data science/analytics/AI methodologies implemented so far in e-health applications. This is achieved by identifying the main techniques and examples available in the scientific literature. A discussion on the benefits these techniques offer to the e-health sector, including a series of best practices, is another contribution of this paper. Finally, we also propose some open challenges yet to be fully explored. There are some recent reviews on related topics, such as those by Matheny, Whicher and Israni (2020) or Rong et al. (2020), among others. Our work contributes to this field by providing a holistic overview regarding the use of data science/analytics methods and AI algorithms in e-health and a discussion of the main open challenges. In our view, this can be equally useful for managers and researchers in the area.

The remaining of this paper is organized as follows: Section 2 offers a description of the primary needs of healthcare organizations and, in general, healthcare systems, as well as a text mining analysis that aims at identifying the most relevant keywords and hot topics. Section 3 provides an overview of the leading data science/analytics methods that can be applied in the e-health context. Section 4 performs a similar analysis in the case of AI algorithms. Several examples of real-life applications are analyzed in Section 5, while the main identified challenges and open research lines are discussed in Section 6. Finally, the main findings of this study are summarized in Section 7.

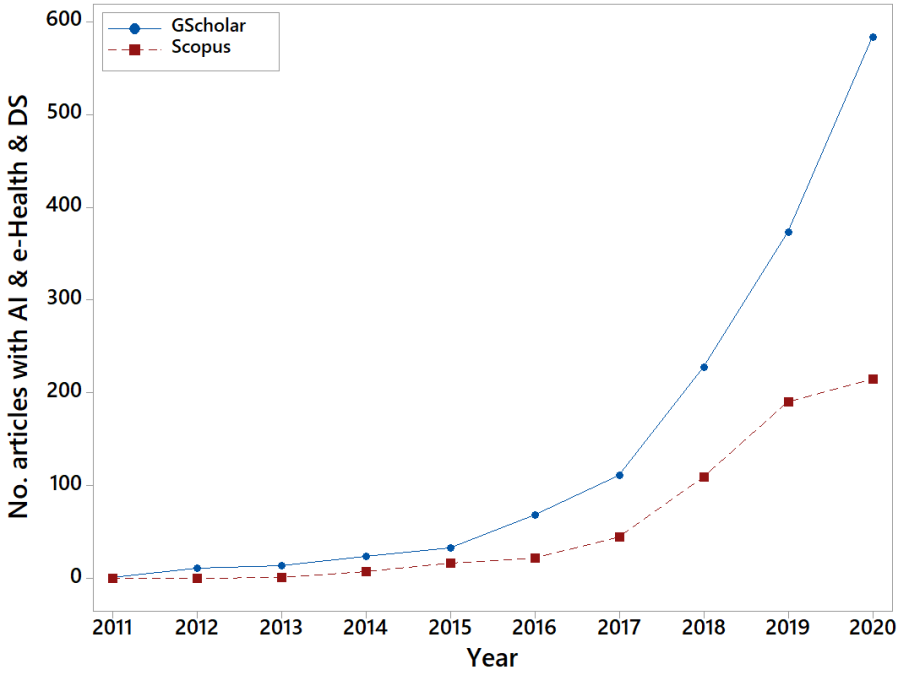

Figure 1. *Number of articles in Google Scholar and Scopus including all the terms: "artificial intelligence", "e-health", and "data science".*

2. Data science/analytics, AI, and current trends in healthcare

Healthcare is one of the biggest and fastest-growing industries in the world. Healthcare management has been changing from disease and volume-focused to patient and value-centered delivery systems in recent years (Huang et al., 2015). The efficient management, analysis, and use of big healthcare data are crucial for providing patient and value-centered care. Most common traditional data management protocols-which are currently in use in healthcare centers-cannot analyze big data efficiently since the complexity and volume of data in healthcare have significantly increased over the past three decades. Therefore, there is an ever-increasing demand for new innovative methods and tools of big data management to support the healthcare industry (Feldman, Davis and Chawla, 2015).

In the last year alone, several alliances have been taking place among some of the biggest pharmaceutical companies and the technology giants in AI use. For instance, Boehringer Ingelheim has announced an agreement with Google Quantum AI to support the research and application of use cases for quantum computing in pharmaceutical research and molecular dynamics simulations (Reinig, 2021). Novartis and Microsoft have reaffirmed their commitment to using AI for drug research and development (Zuest, 2019). AstraZeneca has announced partnerships with Alibaba to develop smart health

services, screening tests, and AI-assisted diagnostic tools in China (Martuscelli, 2018). Capgemini has announced the signing of a contract with Bayer AG to accelerate its digitization (Connatty, 2019). IBM is close to predicting and diagnosing Alzheimer's, for which it has partnered with Pfizer to develop a model for early detection using AI (Terry, 2020). Almirall and Iktos (a company specialized in AI) have signed an agreement to accelerate the discovery of new drugs (Al Idrus, 2019). In the recent COVID-19 pandemic, a Canadian start-up AI company (BlueDot), which tracks and predicts the spread of infectious diseases, alarmed its customers about the spread of atypical pneumonia that was taking place near a shopping area in Wuhan, China. BlueDot was the first organization in predicting the spread of that disease, even nine days before the World Health Organization released its report about the outbreak of a new coronavirus in China (Bowles, 2020).

All this is reflected in the growing number of scientific articles that have been developed in recent years. The large volume of information available in this area creates the need to use intelligent techniques capable of processing large amounts of text and extracting valuable information from it, such as the topics most addressed. For this purpose, 633 scientific papers indexed in the Scopus database for the terms "artificial intelligence", "e-health", and "data science" are taken as a basis for analysis. We have completed a text mining process on their titles and abstracts. This analysis allows us to: *(i)* automatically identify the most cited keywords in all these papers; and *(ii)* automatically generate a list of hot topics in this research field. The latter goal has been achieved by employing a latent semantic analysis with the non-negative matrix factorization (NMF) algorithm. NMF is a feature extraction algorithm that combines attributes to produce meaningful topics. It decomposes multivariate data by creating a user-defined number of features, each of which is a linear combination of the original attribute set (Huang, Zhou and Zhang, 2012b). This algorithm enables the modeling of the topics, i.e., to extract the significant topics that recur in the corpus or media group of similar documents. Documents are decomposed into topics and topics into words. This technique requires that, in the case of texts, all documents to be analyzed have a similar length. We analyze abstracts and titles, which generally have a similar length in terms of number of words, so this requirement is met.

At a technical level, the NMF decomposes the matrix of visible variable (V), which is the input, into two smaller matrices, the document-topic matrix (W) and the topic-term matrix (H). Matrix V contains a count of the occurrence of each word (document by term frequency), matrix W, presents for each row one document by the non-normalized probabilities of topics. The W matrix presents for each row one document per non-standardized probabilities of the topics, and allows us to interpret that two terms that appear together frequently form a topic and each term gives more contextual meaning. Matrix H allows to establish the number of topics which are interpreted as every two terms appearing together frequently form a topic. Each term gives more contextual meaning to the term with which it is grouped, and if a term appears frequently in two themes, they are likely to be related. This allows us to establish the number of topics

(t) to determine the size of these matrices. In addition, it has the advantage that each topic is interpretable, which is not the case with other matrix decomposition methods such as principal component analysis and vector quantization that only use non-negative numbers (Snasel et al., 2007).

Both the data preprocessing and the implementation of the NMF algorithm have been performed with a Python script. Data preprocessing consists of the cleaning necessary before applying the algorithm. For this purpose, in the Python script an exclusion list is constructed with the words to be removed, such as articles, connectors, prepositions and determiners found in the data set. The "counter" function of the "collections" library is used, which is a container that counts the number of times that there is an equivalent value in the dataset to be analyzed. This allows us to know the number of words not significant for the analysis that are still in the input data of the algorithm. The cleaning process is repeated until the non-relevant words and their equivalents are not found in the dataset. The topic extraction model is with the NMF decomposition algorithm in the scikit-learn library (Pedregosa et al., 2011) as described by Liu (2017). The parameters implemented were $n\_components = 5$ and $n\_top\_words = 10$. Therefore, the results provide the 10 most popular words and the top 5 topics in the analyzed articles. The Python code is available at https://github.com/NMF_ehealth.git As shown in Figure 2, the most popular words in the set of analyzed articles are: "data", "health/healthcare", "research", "IoT", "security", "smart", "information", "system", and "learning".

Likewise, by configuring the NMF algorithm (Python scikit-learn version) to determine the top five topics, we obtain the results presented in Table 1. These topics are determined from organizing the sets of five keywords generated by the algorithm. Articles, connectors, prepositions and determiners are added to make sense of the five keywords in each set. These topics can be considered as some of the most popular emerging research lines in the literature, and they include keywords such as "machine learning", "IoT", "security", "blockchain", "privacy", "big data", "medical analytics", "cloud and fog computing", etc.

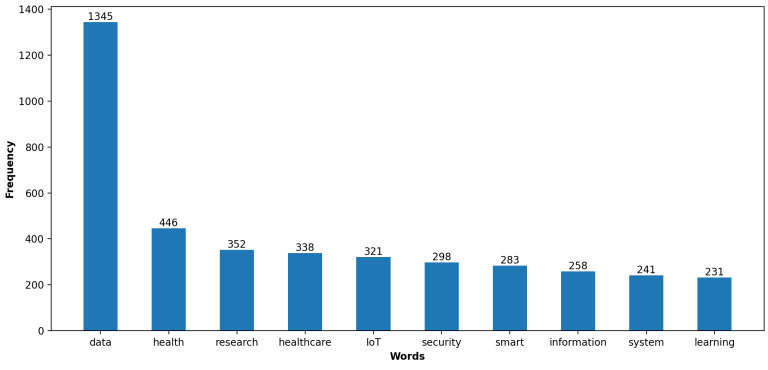

Figure 2. *Most common words in the articles indexed in Scopus including all the terms: "artificial intelligence", "e-health", and "data science".*

Table 1. *Topic modeling results obtained with the NMF algorithm for the articles indexed in Scopus including all the terms: "artificial intelligence", "e-health", and "data science".*

No.	Five-word set	Topic	References
1	learning, machine, deep, model, recognition	machine & deep learning for recognition	Shatte, Hutchinson and Teague (2019) Yu, Beam and Kohane (2018) Kavakiotis et al. (2017)
2	IoT, internet, things, devices, security	IoT devices security	Al-Garadi et al. (2020) Din et al. (2019) Makhdoom et al. (2018)
3	blockchain, technology, applications, consensus, research	blockchain technology in security and privacy applications	Chukwu and Garg (2020) Roy et al. (2018)
4	data, big, analytics, processing, medical	big data in healthcare medical analytics	Wang and Alexander (2020) Syed et al. (2019)
5	access, control, encryption, data, attribute	cloud and fog computing for privacy and security	Sun (2020); Dang et al. (2019) Mutlag et al. (2019) Puliafito et al. (2019)

A second text mining process on titles and abstracts is performed only for the terms "e-health" and "artificial intelligence" in the Scopus database, considering only articles, literature reviews, conference papers and book chapters. 403 documents are analysed using the same parameters as described above. Discounting the words "health" with 487 occurrences, "data" with 397, "medical" with 252, and "healthcare" with 247, we obtain the 10 most common words among the analysed documents shown in Figure 3. The most common words in the titles and abstracts of these documents are: "information", "patients", "systems", "e-health", "decision", "artificial", "clinical", "learning", "monitoring", and "smart". This reveals a trend towards AI research in healthcare related to patient data and clinical information, as well as monitoring, and decision-making. We have also extracted the top five topics using the NMF model. Table 2 presents the results showing that the most emerging lines of AI research relate to medical patient information, decision-making, IoT and cloud systems integration, the COVID-19 pandemic, and the use of machine learning in the area of diseases. As in the previous text mining process, these five topics are the result of the interpretation of the resulting set of five keywords generated by the algorithm.

3. Data science/analytics methods in e-health

The emergence of e-health has created a significant demand for data analysis of people's health and administrative healthcare processes. For this reason, this section presents an

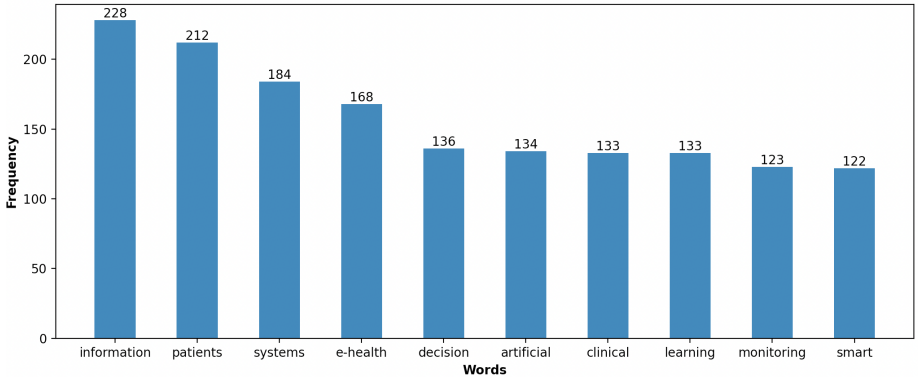

Figure 3. *Most common words in the titles and abstracts of indexed documents in Scopus including all the terms: "artificial intelligence" and "e-health".*

analysis of the methodologies and techniques of data science/analytics implemented in e-health systems, also considering AI methodologies. Data science allows us to apply both quantitative and qualitative methods to solve significant problems and predict results (Waller and Fawcett, 2013). Big data and data analysis go hand in hand because data is considered the raw material of data science (Larson and Chang, 2016). When applying these methods in e-health, one of the main goals is to extract knowledge from data in order to improve patient care (McIntosh et al., 2016). The data collected from software applications in e-health, apart from being available in large amounts, are also characterized by not having well-defined structures and being heterogeneous. Therefore, they require methods that order, filter, process, and extract patterns from large amounts of data in order to make them valuable. According to Van Der Aalst (2016), data science refers to data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, all types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects. Hernán, Hsu and Healy (2019) organize data science into three tasks: (i) description, i.e., referring to quantitative analysis ranging from elementary calculations to unsupervised learning algorithms and intelligent data visualization techniques; (ii) event prediction, i.e., using elementary predictive calculations, such as correlation of variables, and methods for recognizing patterns in data and supervised learning algorithms; and (iii) counterfactual prediction, i.e., using data to predict events in different scenarios involving causal inference. Table 3 summarizes the data science methods applied in different e-health contexts. It also mentions the general techniques with some possible applications.

According to Galetsi and Katsaliaki (2020), Descriptive Analytics techniques are used to identify problems and trends in the data. Within the examples of applications in e-health, they are mainly used to identify diseases or provide analysis on medical information obtained from different sources, such as mobile applications or medical test equipment. Wang and Hajli (2017) highlight the five main business capabilities of big

Table 2. *Topic modeling results obtained with the NMF algorithm analyzing the titles and abstracts of indexed documents in Scopus including all the terms: "artificial intelligence" and "e-health".*

No.	Five-word set	Topic	References
1	health, information, data, care, patients	patients health care information data	Mamdouh et al. (2020) Susanto (2017) Sethia et al. (2016)
2	decision, support, clinical, based, making	clinical-based decision-making support	Hu et al. (2016) Impedovo, Pirlo and Vessio (2018) Aldape-Pérez et al. (2018)
3	IoT, applications, things, internet, cloud	IoT / internet cloud applications	Zhu et al. (2015) Miori and Russo (2012) Lakshmanaprabu et al. (2019) Ruiz-Zafra et al. (2013)
4	covid, 19, pandemic, monitoring, detection	pandemic monitoring detection COVID-19	Channa, Popescu and Malik (2020) Chakkor et al. (2021) Lagos-Ortiz et al. (2020)
5	learning, machine, disease, data, using	machine learning using disease data	Kavakiotis et al. (2017) Pereira et al. (2019) Ferreri et al. (2018)

data analytics in healthcare, which are: (i) traceability in patient monitoring, including lab results, medication, historical data, and current status; (ii) analysis of structured and unstructured data, e.g., comparative analysis of images, voices, texts, etc; (iii) speed-up decision-making with automatic notifications or visual reporting; (iv) interoperability with the integration of heterogeneous data from different sources; and (v) prediction of patients behavior. These capabilities not only exemplify the possible applications of descriptive techniques in healthcare, but also point to the power they have to improve healthcare processes. For some authors, the fact that prediction is among the capabilities of descriptive analytics shows how the fields of data science and analytics are connected. Predictive analytics is the field of analytics in which future events are foreseen (Mishra and Silakari, 2012). As stated by Van Calster et al. (2019), the large investments in new AI technology reflects the value of this field to healthcare. This is due to its ability to diagnose individuals most likely to suffer from a disease, or to predict the evolution of diseases or viruses. This helps to make decisions in the treatment of patients, increasing the probability of recovery. It is also used in administrative issues to manage resources in seasons with a higher or lower probability of a high flow of patients in medical centers. From a general perspective, predictive techniques in e-health systems can make processes in the healthcare field more efficient.

Table 3. *Data science methods applied in e-health.*

Data science fields	Techniques	Examples of e-health applications	References
Descriptive analytics	Data retrieval & collection (statistics & data visualization)	Patients diagnostic, epidemic recognition, patients management with mobile apps, visual analysis tools for clinical data, reporting systems, social media analysis, and data warehouse tools.	Galetsi and Katsaliaki (2020) Wang and Hajli (2017)
Predictive analytics	Machine learning, probabilistic modeling, & statistical analysis	Disease prediction, software as a service, medical decision support system, patient flow and use of medical resource prediction.	Lepenioti et al. (2020) Van Calster et al. (2019) Mishra and Silakari (2012)
Prescriptive analytics	Logic-based modeling, evolutionary computation, mathematical programming, reinforcement learning, & simulation	Decision-making automation, scheduling problems, balanced assignment workload and resources, health management assessment, and cloud computing.	Javaid et al. (2021) Nandankar et al. (2021) Shah, Bhat and Khan (2021) Wijnhoven (2021) Bertsimas and Kallus (2020) Lepenioti et al. (2020)

There is also Prescriptive analytics, which is a field that prescribes optimal decisions for operations research (OR) and management science, together with machine learning techniques (Bertsimas and Kallus, 2020). In general, these techniques provide meaningful insights and support decision-making for organizations, thus giving them competitive advantages. This field is considered the evolution of the descriptive and predictive fields of data science, due to the capability of offering intelligent recommendations based on the analyzed and predicted data (Lepenioti et al., 2020). Since this is a field that is still in its infancy (Lepenioti et al., 2020), there are not too many works that integrate these data-driven techniques into e-health applications. Among the 25 papers available in the Google Scholar database for the first semester of 2021 and indexed under the terms "prescriptive analytics" and "e-health", some interesting applications are: *(i)* healthcare monitoring systems integrating a cloud IoT (Shah et al., 2021); *(ii)* a groundbreaking analysis system for large-scale healthcare data, which allows the use of fog computing and cloud systems to deal with data processing, storage, and classification problems (Nandankar et al., 2021); *(iii)* a dental 4.0 decision-support system to provide users with high-quality and personalized experiences (Javaid et al., 2021); and *(iv)* the implementation of an analytic clinical decision support system that analyzes medical data to predict the probability of sepsis in prematurely born infants, which can support physician decision-making on antibiotic stewardship (Wijnhoven, 2021).

With the goal of identifying some recent applications of data science methodologies in e-health, a similar study was carried out using the Scopus database. A total of 10 new papers were found for the first semester of 2021. These papers combined the terms "artificial intelligence", "e-health", and "data science" in the title, abstract, or keywords sections. Table 4 summarizes the most recent application fields of these techniques in

Table 4. *Identified data science applications in e-health within the literature available in Scopus database.*

Data science techniques	Application field
Data retrieval & collection (statistics & data visualization)	Epidemiological analysis (Pfeiffer and Stevens, 2015) Computer-assisted surgical skills (Vedula, Ishii and Hager, 2017) Preventive health management system (Neubert et al., 2019) Mental healthcare (Naslund et al., 2019) Correlations between clinical medicine subjects (Chen et al., 2020b)
Machine learning	Automatic behavior identification (Crocamo et al., 2020) Behavior assessment (Liang et al., 2020) Symptom classification (cardiology) (Oliver et al., 2018; Spanakis et al., 2017)
Modeling & simulation	Virus propagation risk analysis (Chatterjee, Gerdes and Martinez, 2020) Disease progression (Idrees and Sohail, 2021) Optimize emergency departments operations (Vanbrabant et al., 2019) Human organs simulation (Serra et al., 2021; Quarteroni et al., 2017)

e-health. Some of the applications that have been identified are those referring to the development of behavioral assessment systems (Spanakis et al., 2017; Crocamo et al., 2020) and to data management systems (Neubert et al., 2019), where data collection methods are implemented through different techniques. Spanakis et al. (2017) develop an adaptive feedback module of an e-coach application on eating behaviors. The authors use the acquisition of ecological momentary assessment (EMA) data through a mobile application. EMA is a set of methods that assess research subjects in their natural environment, in their current or recent states, at predetermined events of interest, and repeatedly over time (Moskowitz and Young, 2006). It allows data to be collected online and in real-time, thus generating more accurate and valuable results. As it is based on questionnaires, it reduces recall bias. It is a methodology that can easily be extended to include information from sensors, e.g.: assessing stress levels or GPS information to evaluate energy expenditure information (Spanakis et al., 2017). In addition, Crocamo et al. (2020) implement an automatic system that identifies keywords or *#hashtags* in Twitter publications that cite alcohol-related behaviors. They work with a systematic tracking process on Twitter using an ad-hoc Python script. After collecting the tweets, they classify all the information to filter out and identify the genuine users. Then, they implement an additional classifier focusing on the linguistic characteristics of the content. All data collection and processing methods are performed with different Python scripts based on the natural language toolkit framework and the scikit-learn library. For data management, Neubert et al. (2019) propose a decentralized system for preventive health based on multi-sensorial fusion (different devices), including heterogeneous data. They develop a mobile application that is the central data node for individual client monitoring, and also contains a data preparation system (data consolidation and synchronization, pre-processing, data selection, and reformatting of data sets) to be stored in the cloud in the required format. The data transfer between the mobile systems and the cloud is based on the formatting of the data to an arranged JavaScript object notation protocol.

In addition to data extraction, cleaning, and transformation, data visualization methodologies have also been applied in the area of e-health in order to extract essential and useful information in support of decision-making. An example of this is the "surgical data science" system. Vedula et al. (2017) conduct a literature review on computer-assisted objective technical capability assessment systems that allow for scalable and accurate assessment of surgeons. They also provide visualized feedback and automated training, thus improving surgical training and maintenance of surgical skills. These systems use data science techniques, such as summary features to represent data, time-series data representations, dictionaries or histogram-based representations, classification algorithms, etc. In more general areas of medical research, authors such as Chen et al. (2020b) explore and comprehensively compile topics from the clinical medicine literature by manually labeling the topics of diseases. In addition to searching in scientific databases and cleaning and integrating data in Microsoft Excel, they apply visualization techniques based on descriptive statistics (e.g., histograms, spatial distribution, etc.) developed in the R statistical software. The results allow them to reach interesting conclusions regarding the relationship to data-driven studies in the field of medicine and health. Pfeiffer and Stevens (2015) provide a literature review on the analysis of temporal and spatial data to support the management of complex animal health problems. According to their conclusions, the opportunity offered by digital technologies in animal and human health requires an interdisciplinary approach that, in addition to the various health areas, also includes information technology.

Data mining is a sub-discipline of data science characterized by discovering knowledge in large databases or extracting patterns from them (Kriegel et al., 2007). It lies at the interface of database technology, pattern recognition, machine learning, and other areas (Hand, 1998). Data mining techniques are also implemented in some areas of e-health, such as classification and clustering algorithms for human behaviors and heartbeat categories analysis (Liang et al., 2020). Artificial neural networks (ANN) are applied for heterogeneous and multidimensional data analysis in general healthcare and psychological interventions (Oliver et al., 2018), respectively. Data mining makes it possible to uncover patterns or trends in human behavior and illness trajectories that were previously not visible. Still, there is a need for large amounts of data in different healthcare areas, such as in mental health (Naslund et al., 2019) or surgical skills for assessment frameworks (Vedula et al., 2017).

The few applications available today in the literature show how implementing these methodologies supports several healthcare processes, ranging from physicians' training and evaluation systems to the use of diagnostic and follow-up systems, including different administrative processes. Furthermore, the number of articles identified indicates that the application of data science methods and the adoption of the terms e-health and data science together have not been widely used until now. The low number of articles explicitly including the term "data science" in their description reflects that the application of these methodologies in e-health is at an early stage yet. Furthermore, many of the articles are literary reviews seeking to establish theoretical foundations and to

identify new research lines related to the applications in e-health, rather than the actual development of methodologies, case studies, and applications.

Table 5. *Number of papers per year by combining the term "e-health" with different AI subfields.*

subfield	Year																			Total
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	
Deep learning	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	8	17	20	34	85
Reinforcement learning	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	2	3	3	5	15
Clustering	0	0	0	1	0	2	0	0	5	0	1	4	3	5	11	9	8	12	5	8	74
Data visualization	0	0	1	0	1	2	0	1	3	5	2	2	1	4	1	2	3	9	3	5	45
Artificial neural network	0	0	0	0	1	2	1	1	1	0	3	2	3	0	2	2	4	5	3	10	40
Natural language processing	1	0	0	0	0	0	1	1	1	0	2	4	4	3	3	2	3	9	4	5	43
Fuzzy logic	0	0	0	1	1	0	1	1	2	0	5	1	4	2	6	4	2	1	2	10	43
Bayesian networks	0	0	0	0	0	1	4	0	0	0	9	1	3	1	2	0	2	1	3	3	30

4. Artificial intelligence algorithms in e-health

As shown in Figure 1, the interest in AI applications to e-health has been quickly rising during the last decade. Among the several fields that can benefit from AI, the healthcare community shows a particular interest due to the amount of data and information that new technologies can provide. As pointed out by Oke (2008), inside the AI field we can find subfields that should be separately considered: machine learning, fuzzy logic, artificial life, data mining, Bayesian networks, knowledge engineering, ANN, reactive systems, semantic networks, computational language, natural language processing (NLP), etc. In particular, machine learning and data mining are subfields that in themselves encompass many subfields of AI, such as deep learning, reinforcement learning, clustering, data visualisation, etc. Data mining can extract usable information from immense raw data sources, primarily as a solid input for subsequent advanced downstream data analyses processes. On the other hand, machine learning has shown to have considerably broad applications in healthcare. Considering the above, Table 5 shows the evolution of the number of papers over the last two decades for the most popular subfields. The number of articles have been extracted from the Scopus database. Deep learning has been the most frequently employed approach in the last years, followed by clustering, data visualization, NLP, ANN, Bayesian networks, and reinforcement learning.

Thus, all these subfields lead to innovations and discoveries in all aspects of medicine, which may have persuaded researchers to focus more on exploring the potential of machine learning in healthcare research, such as the implementation on deep learning compared to data mining techniques like clustering and data visualization. However, these subfields are not worked on separately. Figure 4 presents the number of papers in which techniques from the different subfields analysed have been combined. This information is constructed after a combined search of the Scopus database for "e-health" and two of the subfields analysed. According to the figure, the subfields that have so far been studied together are clustering with all but deep learning and reinforcement learning. These last two subfields are worked together, with deep learning being the only one that

is related to ANN, Bayesian networks, and NLP. Fuzzy logic also seems to be of great interest when working with ANN, clustering, and Bayesian networks.

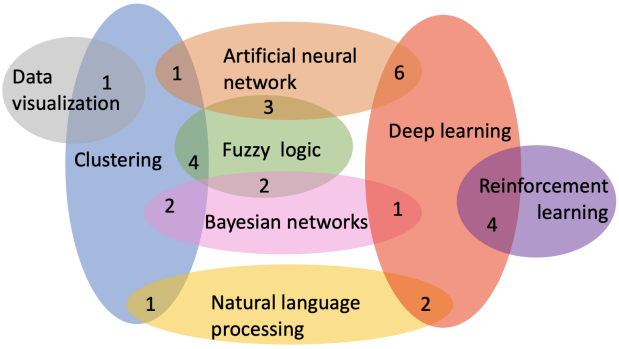

Figure 4. *Number of papers by combining the term "e-health" with two different AI subfields indexed in the Scopus database.*

This section will provide an overview of some of the aforementioned methodologies, as well as a discussion on their applications in e-health. The methodologies have been selected based on the number of publications during the last years. For a deeper discussion on how these methods are employed, the reader is referred to the works included in Table 6.

4.1. Machine learning

The growing use of devices that collect health data (e.g., wearables) favors the development of both, supervised and unsupervised methods, in diagnostics and diseases prediction. These devices allow collecting health data such as heart rate, number of steps, calories, sugar levels, hours of sleep, or images. All of these data can serve as inputs for the machine learning algorithms.

The main objective of algorithm development in the scientific community is to make computers completely autonomous in predicting results. However, in the health field, autonomous machine learning is far from being implemented. Consequently, the integration of a human expert in the circuit allows for better approaches in the complex field of e-health. The interactive machine learning approach can be defined as algorithms that can interact with human or computational agents to optimize the learning process (Holzinger, 2016). In e-health, the algorithms used in the interactive approach can be of particular interest in problems where the lack of data to train models or decision-making can be replaced by the help of the medical experts, where their knowledge and experience can be of significant contribution in solving the problem that would otherwise remain too complex. Some examples where this approach is used are described by Hund et al. (2015) and Lathrop (1994), where subspace clustering and protein folding problems are presented, respectively.

In contrast with the benefits mentioned, interactive machine learning approaches present an open question about their robustness: their evaluation is not only more diffi-

Table 6. *AI surveys by subfields.*

AI field	Topic	Reference
Machine learning	Summarizes a variety of machine learning research techniques in health informatics.	Dua, Acharya and Dua (2014)
	Studies the concept of interpretability in the artificial intelligence field in healthcare.	Al-Garadi et al. (2020)
	Reviews some ML algorithms used for developing efficient decision support for healthcare applications.	Shailaja, Seetharamulu and Jabbar (2018)
Reinforcement learning	Surveys of applications in healthcare, focusing on the discovery of new treatments, personalizing existing ones, and automated medical diagnosis.	Yu et al. (2021)
		Coronato et al. (2020)
		Gottesman et al. (2019)
Data mining	Reviews the utility of various DM techniques like regression, clustering, association, classification in healthcare.	Tomar and Agarwal (2013)
		Koh and Tan (2011)
		Jothi, Rashid and Husain (2015)
		Birnbaum (2004)
Artificial neural networks	Applications of ANN to health care organizational decision-making.	Lisboa and Taktak (2006)
		Shahid, Rappon and Berta (2019)
Natural language processing	Summarizes the applications, techniques, principal challenges of NLP in healthcare.	Friedman and Elhadad (2014)
	Discusses the main challenges of NLP in the healthcare sector.	Carrell et al. (2017)
	Reviews the NLP techniques used in healthcare, their applications and limitations.	Iroju and Olaleke (2015)
	Presents a system that combines both text mining and NLP.	Popowich (2005)
	Focuses on development and advances of NLP methods for clinical decision support.	Demner-Fushman, Chapman and McDonald (2009)

cult, but also these methods are not easy to replicate. The reason is that the contributions provided by human agents are subjective and cannot be easily imitated. Some applications of machine learning to diseases prediction can be found in Senders et al. (2018), Goldstein, Navar and Carter (2017), Weng et al. (2017), Churpek et al. (2016), Taylor et al. (2016), Kruppa et al. (2014), and Singal et al. (2013).

Reinforcement learning is a powerful and increasingly popular subfield of machine learning, which studies optimal sequential decision-making under uncertainty. This approach is highly relevant in the fields of dose optimization (Hrinivich and Lee, 2020; Tejedor, Woldaregay and Godtliebsen, 2020; Watts et al., 2020) and robotic-assisted surgery (Gao et al., 2020; Pore et al., 2021; Su, Huang and Hannaford, 2021). There are also other recent works on public health (Weltz, Volfovsky and Laber, 2022; Kwak, Ling and Hui, 2021), surgical decision-making (Datta et al., 2021), physical activity mobile health application (Liao et al., 2020), and cancer detection (Liu et al., 2019), among others.

4.2. Data mining

The application of data mining (Fayyad, Piatetsky-Shapiro and Smyth, 1996) can provide knowledge useful to support clinical decision-making. In the healthcare and biomedical fields, data mining deals with privacy and legal issues, as well as on improving the quality of data available in these disciplines. Paper-based or scanned-digital formats, heterogeneity of hospital information systems, or different lab tests for the same disease constitute big challenges for researchers. Here is where data mining can help by providing tools to extract information from large and unstructured data sources.

Data mining algorithms can be classified into two categories: descriptive and predictive (Tamilselvi and Kalaiselvi, 2013). The first one aggregates records with similarities, thus discovering unknown relationships in data. The second one deduces prediction rules from training data and applies these rules to unpredicted data.

4.3. Artificial neural networks and natural language processing

ANN are computer programs that try to reproduce how the human brain processes information, i.e.: learning through experience, by recognizing patterns and relationships in data. An ANN is constituted by hundreds of single units (artificial neurons), which have weighted inputs and one output. In healthcare, deep neural networks face big challenges that need to be addressed before they can be used in the daily life of patients. They require an exceptionally huge amount of information to function better than other strategies. They are extremely costly to train due to complex information models. Besides, deep learning needs costly GPUs and hundreds of machines. This enhances its costs of operation. There is not enough standard theory to direct people in choosing proper deep learning resources because it needs information on topology, preparing strategy, and other necessities. Therefore, it is hard to be used by less knowledgeable end users (Kim et al., 2020).

When managing clinical reports, one of the biggest challenges for analysts is that information is not structured. In fact, it is usually written in natural language as plain text. Hence, NLP techniques are responsible for extracting knowledge from this unstructured data, analyzing the information it contains, and providing it in a format that the electronic healthcare systems can easily understand. More than twenty years ago, Friedman and Hripcsak (1999) and Baud, Rassinoux and Scherrer (1992) had already highlighted the importance of NLP in medicine. Friedman and Elhadad (2014) ranks the top challenges regarding NLP and also concludes that software-generated recommendations should always be supervised by human decisions.

4.4. Use of AI in e-health

Many studies have demonstrated the wide applications of AI in e-health. As stated in Jiang et al. (2017), AI "is bringing a paradigm shift to healthcare". These authors provide a literature review on AI applications in healthcare, including areas such as cancer, neurology, and cardiology. A similar review is provided by Yu et al. (2018), who support

the idea that AI has contributed to improving diagnosis and decision-making in healthcare, while recognize its potential in fast disease detection and customized treatments. Authors such as Davenport and Kalakota (2019) or Emanuel and Wachter (2019) highlight the possibilities that AI offers in areas such as automatic examination of radiology and pathology images. However, they suggest that the main challenge AI has to face in healthcare is not a technological one, but a cultural one –so that these techniques are adopted and employed in daily clinical practice. He et al. (2019) identify the main challenges regarding the practical implementation of AI into daily healthcare practice. Among these issues, they include data sharing and privacy as well as algorithms' transparency. According to Reddy, Fox and Purohit (2019), patient administration, clinical decision support, patient monitoring, and healthcare interventions are the healthcare areas in which the use of AI can be more beneficial.

5. Data science & AI best practices in e-health

Most e-health tools not only include the use of Internet-based applications, but also products, systems, and services. Health portals, telemedicine services, electronic health records, or health information networks are just some examples. As already pointed out in several works (Widmer et al., 2015; Triantafyllidis et al., 2015; Warmerdam et al., 2010), AI methods have an enormous potential for improving healthcare, reducing costs, and developing smart digital health interventions. Such enhanced interventions can lead to remarkable outcomes, both for patients and healthcare providers (Murray et al., 2016; Obermeyer and Emanuel, 2016). This section presents relevant application domains related to the application areas identified in Section 2. It is structured in five subsections: patient care, public health, healthcare management, COVID-19, and other topics.

5.1. Patient care

Regarding patient care, we identify the following areas of applications:

- *Diagnosis:* A large number of works propose an AI methodology to either perform or assist the expert with the diagnosis. For instance, Kermany et al. (2018) create a diagnostic tool based on a deep-learning approach for screening retinal and lung pathologies for age-related macular degeneration and pediatric pneumonia, respectively. The diagnostic results of the training that neural network framework was comparable to that of health professionals. Indeed, that AI-based diagnostic method was even able to provide more detailed diagnoses, especially at the onset of pathologies when the clinical symptoms might not be apparent to healthcare providers. Similarly, Guo et al. (2020) propose a deep learning model for real-time automated diagnosis of precancerous lesions and also to assist the diagnosis of esophageal cancer. The model has high sensitivity and specificity for both endoscopic images and videos. Additionally, NLP has been used in the diagnosis of diseases such as Aphasia, a neurological disorder (Rao and Venkatesh, 2021). It

is also applied in medical decision support systems for the diagnosis of patients. This is performed by semantic analysis of their available medical records (Amato et al., 2018; Iram and Gill, 2018). Other solutions are focused on bringing the healthcare sector online. Virkar et al. (2021) develop a system that provides information on remote doctor connection and online identification and treatment, and describes the prediction of various diseases using a web application as an interface to store the data in a database. The NLP collects the necessary symptoms and sends them to the doctors. This is used to predict the exact disease the patient may have using the Random Forest algorithm. Thus, utilizing such AI-based frameworks can lead to earlier and more accurate diagnoses, which are crucial for curving the progression of diseases and selecting the most effective treatments for patients.

- *Personalized medications and diets*: The traditional "one-size-fits-all" approach is inefficient in many scenarios. AI is increasingly playing a pivotal role in personalized medications. An example is related to Antimicrobial resistance (AMR), a phenomenon in which microorganisms resist antimicrobial drugs. AMR poses a serious risk to preventing and treating infections caused by viruses, bacteria, and fungi. AMR will lead to nearly 10 million death and 100 billion dollars economic damage every year by 2050 (O'Neill, 2014). AI-based predictive models can be applied against AMR. These models can be used in predicting AMR and suggesting the best dose and time of antimicrobial treatment, as well as the best combinations of antimicrobial peptides and antibiotics for each patient (Lv, Deng and Zhang, 2020). Zeevi et al. (2015) use factors –such as gut microbiota, anthropometrics, blood parameters, and dietary habits– to create a gradient-based boosting regression for predicting the post-meal glycemic fluctuations in real life in a randomized controlled study. The results showed that the personalized diet created according to the post-meal glycemic simulation significantly optimized the post-meal glucose level in the study individuals. Such approaches can be used in various nutritional interventions in different diet-related medical problems, such as diabetes, obesity, and nonalcoholic fatty liver disease.

- *Personalized care:* AI also contributes to personalize care and interventions. For instance, Barrett et al. (2019) perform a study in which a "virtual doctor" is designed relying on AI, serious gaming, and patient coaching. This doctor boosts advanced and personalized self-care for heart failure patients, where the patients themselves perform standard care tasks. Burns et al. (2011) use a mobile-based, multi-component type of intervention that worked based on some machine learning models in order to predict the emotions, activities, mood, motivational and cognitive states –as well as social and environmental changes– for some patients who were suffering from depression. That intervention also enabled the patients to have access to the coaching feedback graphs provided by their caregivers. At the end of the study, nearly 90% of the patients indicated that the machine learning-

based mobile intervention helped them treat their depression and control their mental symptoms, and they were satisfied with that.

- *Treatment optimization:* Planning a treatment tend to be a challenging and key task. Many authors have proposed AI-based applications to enhance this process. For instance, Wang et al. (2019) review AI-based applications for radiotherapy treatment planning aiming to minimize the normal tissue damage while persevering sufficient tumor control. Similarly, Cui et al. (2021) explore the use of random survival forest models to predict optimal regimen classes for individual patients and each line of therapy relying on baseline characteristics. The patients are adult females with HR+/HER2- breast cancer and the aim of the models is to maximize overall survival and time to treatment discontinuation based on electronic health records. Shamir et al. (2015) propose a clinical decision support systems based on machine learning clinical decision support systems based on machine learning (support vector machines, naïve Bayes, and random forest are considered) to optimize combined stimulation and medication therapies for Parkinson's disease.

- *Assisted or automated prescription:* In the context of a growing amount of available clinical data and an increasing focus on personalized care, assisted or automated prescription constitutes one of the most promising AI applications. For instance, Blansit et al. (2019) propose an approachable to prescribe imaging planes for cardiac MRI based on deep learning, which relies on the localization of anatomic landmarks. Barbieri et al. (2019) train a model to manage blood pressure, fluid volume, and dialysis dose in end-stage kidney disease patients. According to the authors, these models can help to anticipate patients' reactions through simulation, which can help choose the best treatment for each patient. Leaflets contain information on the administration of medicines, their composition, warnings or precautions for use that is difficult for patients to understand. Dascalu et al. (2019) have designed an intelligent platform based on NLP techniques for drug administration. Its functionalities include adding specific drugs to the user's profile, searching for possible contraindications or side effects and defining drug administration alerts on schedule, based on the doctor's prescription. This enables improved healthcare for users and ensures self-education.

- *Triage:* Determining the priority of patients' treatments by their condition or likelihood of recovery is critical in hospitals, especially during a crisis. For instance, Kim et al. (2018) propose a classification model for survival prediction to perform a precise triage. The authors compare the performance of different methodologies, for example, logistic regression, random forest, and deep neural networks, and design a consciousness index capable of remote monitoring through wearable devices.

- *Surgery:* According to Hashimoto et al. (2018), AI has the potential to revolutionize the way surgery is taught and practiced. The authors review machine

learning applications, ANN, NLP, and computer vision. They also point out the models' low interpretability and the difficulty to determine causal relationships as the main limitations of AI. Riise, Mannino and Burke (2016) propose a new generalized model for surgery scheduling problems. Zhong et al. (2014) present a decision support system to develop master surgery schedules. Bai, Storer and Tonkay (2017) present a gradient-based algorithm in order to improve some aspects, like the cost incurred from patient, waiting time, blocking time, operating rooms overtime, etc.

- *Pregnancy management:* Management of the pregnancy aims to reduce child and maternal mortality by increasing pregnant women's access to high-quality health services. For instance, Moreira et al. (2019a) design smart mobile-health applications that use machine learning for pregnancy monitoring. These applications are able to predict high-risk situations during gestation. Predicting the risk of postpartum depression during pregnancy through biomedical and socio-demographic data analysis constitutes another task where AI may be potentially useful (Moreira et al., 2019b).

5.2. Public health

Best practices regarding the use of data science and AI in public health have also been documented in recent articles. Some of the most relevant ones are included next:

- *Forecasting demand for emergency department services:* Accurately predicting the demand of services in an emergency department is essential to provide efficient and high-quality services. Traditionally, these predictions have been made using historical data and experts' opinions, but some recent works have attempted to add Internet search data. Ho et al. (2019) analyze search volume data retrieved from Google Trends, and construct regression-based predictive models. The method used is relatively simple –multiple linear regression models–, but represents a useful tool to address congestion problems. Likewise, Martin et al. (2012) use the information from patients' phone calls to develop a model to alert health professionals that could predict when patients with critical conditions, such as those who suffered from cardiovascular and lung diseases, who also did not have appointments would go to the medical center for treatment. When his model was deployed over six months, it was able to predict nearly 70% of the unplanned events, which helped give more time to the health system to manage their resources better for admitting those patients. A number of experts have been working in this subject, developing forecasting strategies to preview the arrivals to emergency department (Billings et al., 2013; Kam, Sung and Park, 2010; Hoot and Aronsky, 2008; Hoot et al., 2007).

- *Screening:* In many countries, there are population-based screening programs, with the cancer screening program as one of the most common. These programs

may have restrictions due to budget constraints and lack of experts. Chen et al. (2020a) describe an example of the potential of AI in this context. The authors present a wristband device based on AI to detect atrial fibrillation (AF). AF is a medical condition in which irregular heartbeat can lead to stroke or cardiac arrest. Detection of AF can be challenging as some patients may not have the symptoms while been screened. The wristband was equipped with sensors that measured single-channel electrocardiogram (ECG) and photo-plethysmography (PPG) and an AI algorithm designed to detect AF based on the ECG and PPG input. The reading accuracy, specificity, and sensitivity of wristband were 93%, 96%, and 88% for PPG, and 95%, 99%, and 87% for ECG. Some physicians also evaluated the wristband-recorded ECG. Their accuracy, specificity, and sensitivity of the physicians' judgment were 97%, 98%, and 97%, which were close to those of the wristband algorithm. The convenience of using this method has great potential for long-term screening patients that may suffer from AF, especially in individuals that may have minimal or inconsistent detectable AF symptoms. Another example is described in Bao et al. (2020), which explores an AI-assisted cytology system in a cervical cancer screening program. It improves sensitivity with clinically equivalent specificity.

- *Epidemics:* Data analysis is critical to track outbreaks and design effective strategies to curve epidemics. For example, Ray and Reich (2018) make predictions of infectious disease dynamics with ensemble methods. In particular, the authors predict influenza season timing and severity measures in the United States, both at the national and regional levels. Ganasegeran and Abdulrahman (2020) analyze the role of ineffectively preempting, preventing, and combating the threats of infectious disease epidemics, as well as facilitating the understanding of health-seeking behaviors and public emotions during epidemics. In the biomedical field, text analysis is performed to identify and extract disease symptoms and their associations from biomedical text documents retrieved from the PubMed database using NLP and information extraction techniques to identify feasible disease symptoms (Abulaish et al., 2019).

- *Facing fake news:* The spread of inaccurate information on the internet is a daily occurrence. The impact on people's lives is one of the majors concern in the field of public health. Most of the intelligent techniques tackle this problem are developed mainly using NLP and machine learning (Mesquita et al., 2020). For example, Pulido et al. (2020) conduct an analysis of social media such as Reddit, Facebook and Twitter where they identify that messages focused on false health information are mostly aggressive, those based on evidence of social impact are respectful and transformative. Parfenenko et al. (2020) propose an ANN for the classification of publications on medical care topic into true and misinformative in different WordPress forums to manage this type of publications.

5.3. Healthcare management

We list several types of applications in healthcare management. For each type, a brief description and a reference to a recent example are also provided.

- *Healthcare logistics:* Logistics constitute a strategic function of hospitals' management, senior citizens' rest homes, pharmaceutical companies, etc. The decision-making in this field aims to reduce errors, enhance process quality and reduce waiting times. An increasingly popular topic is home healthcare, which has been boosted thanks to unstoppable technological progress. An example is described in Fikar et al. (2016), where the authors develop a discrete-event driven metaheuristic for dynamic home service routing with synchronized trip sharing. Likewise, in Lostumbo et al. (2021), the authors propose a hybrid method, combining simulation with reliability analysis, to improve supply chains in the healthcare sector. A review of other works related to home healthcare logistics can be found in Fikar and Hirsch (2017), where several applications and different approaches are enumerated.

- *Resource forecasting and optimization:* The increasing demand for resources and the limited capacity in the healthcare sector has increased the use of tools based on forecasting, simulation and optimization for resource management. Some works optimize, for example, the management of hospital beds and the personnel required for each bed modeling the problem as an integer linear programming models, forecast the demand for specialists with ARIMA and linear regression models (Ordu et al., 2021). Others plan patient capacity and patient post-hospitalization fate using decision models based on survival trees (Garg et al., 2012). Ganguly and Nandi (2016) use analysis of variance (ANOVA) to identify drivers of demand, and autoregressive integrated moving average (ARIMA) to develop a forecasting model for optimal healthcare staff scheduling based on patient arrival rates. Similar models have been developed to predict patient visits in the emergency department (Khaldi, El Afia and Chiheb, 2019), and the demand for health diagnostic service such as endoscopy service (Harper, Mustafee and Feeney, 2017). Ellahham and Ellahham (2019) present a review on the applications of artificial intelligence to improve patient management and resource allocation in hospitals.

- *Medicine supply chain network design:* Logistics activities are essential for an efficient and sustainable distribution of medicines. Recently, Goodarzian, Hosseini-Nasab and Fakhrzad (2020) propose a hybrid particle swarm optimization and a genetic algorithm to achieve Pareto solutions for the design of a medicine supply chain network.

- *Treatment/surgery scheduling:* When scheduling surgeries or treatments, such as chemotherapy, a number of objectives and restrictions related to costs, elapsed times, available rooms, experts, and equipment, have to be optimized. For example, Belkhamsa, Jarboui and Masmoudi (2018) present two metaheuristics –an

iterative local search approach and a hybrid genetic algorithm– to address the operating room surgery scheduling, with resource constraints in three stages, namely: preoperative, intraoperative, and postoperative stages. The goal is to minimize the maximum end time of the last activity in stage 3 and the total idle time in the operating rooms. In a different work, Martins et al. (2021) propose a metaheuristic optimization algorithm to support medical staff when assigning and scheduling treatments to cancer patients. The reader interested in healthcare scheduling problems is referred to Abdalkareem et al. (2021), which covers patients' admission scheduling problems, nurse scheduling problems, operation room scheduling problems, surgery scheduling problems, etc.

- *Healthcare facility location-allocation:* Making location-allocation decisions related to healthcare facilities may be a challenge because of multiple conflicting objectives and stakeholders. In this context, Wang, Shi and Gan (2018a) put forward a practicable hierarchical model to characterize the trade-off between social, economic, and environmental factors. The authors present a bi-level multi-objective particle swarm optimization algorithm to deal with the location decision and capacity adjustment.

- *Assessment of the hospital performance:* This constitutes a challenging task because of the high number of related indicators. For instance, Downing et al. (2017) build a semi-supervised machine learning algorithm, which highlights the similarities and differences between hospitals and detects hospital performance patterns for 1 614 U.S. hospitals.

- *Brand management and marketing:* Plenty of strategies in these fields are data-driven. For example, Oztekin (2018) develop data analytic models to help marketing managers identify locations to host peer-to-peer educational events for healthcare professionals.

- *Pricing and risk:* Data has always played an important role in insurance. The increasing amount of available data may improve the process of pricing and risk management. For example, Kshirsagar et al. (2020) assess machine learning models aiming to predict the per member per month cost of employer groups in their next renewal period. The authors conclude that these models may compute an accurate and fair price for health insurance products without losing interpretability.

- *Fraud detection:* Medicare fraud, waste, and abuse cause huge losses, but traditional detection methods tend to be time-consuming and have low accuracy. In recent years, AI has been extensively used for fraud detection, and several works can be found in the field of e-health. It is challenging because of class-imbalance. Zhang and He (2017) propose a method for medicare fraud detection. It consists of two parts: first, a spatial density-based algorithm, called improved local outlier factor; second, a robust regression to analyze the linear dependence between

variables. Johnson and Khoshgoftaar (2019) choose the medicare fraud detection task to compare several deep learning methods built to deal with the class imbalance problem. The authors employ different data-level techniques, such as random over-sampling, random under-sampling, a hybrid one, and several algorithm-level techniques, such as a cost-sensitive loss function focal loss, and the mean false error loss.

- *Patient/user satisfaction:* Patient-oriented interactive tools are a great source of data that provide evidence for strategic planning for e-health development. An analysis of US hospital websites concludes that most hospitals need basic e-commerce tools for their patients/users (Huang and Chang, 2012). A study of more than 200 patients in the USA developed by Huang, Chang and Khurana (2012a), showed that they are interested in access to information such as medical records and lab results. In addition to the patient perspective, it is also noted that there are no ways to measure the success of e-health implementations in serving their patients/users. However, there are some examples of e-health systems that provide information on user/patient satisfaction. Silva et al. (2018) present a satisfaction and usability evaluation of a web-based clinical decision support system called HADA for antenatal care assisting in obstetric risk assessment. This provides a more effective and efficient use of resources and increases the capabilities of professionals and satisfaction for both professionals and patients. Similarly, Lan Hing Ting et al. (2021) study the development of the use of a robotic assistant for geriatric patients. Their results show that the implementation is feasible, as the performance and user satisfaction is promising. Rubrichi, Battistotti and Quaglini (2014) present a system for automatic evaluation of users' perception of the quality of an outpatient visit reminder system based on the short message service (SMS). The automatic interpretation of the content of these messages is useful for monitoring and improving health service performance.

5.4. COVID-19

The COVID-19 pandemic is severely affecting health systems and economies. In this subsection we review some recent examples of AI applications aimed at fighting the COVID-19 pandemic as well as other diseases.

Regarding the molecular aspects of this disease, AI can be used to identify and visualize the molecular structures of the SARS CoV-2 proteins, evaluate different existing drugs, and design new medicines that may help control the disease. Data science methods might also be critical for vaccine development, the design of more accurate diagnosis methods, and increasing our knowledge about the disease's molecular and clinical pathology. For example, Jumper et al. (2020) have designed a model called AlphaFold that predicts the three-dimensional structures of proteins based on their amino acid sequences. AlphaFold has been used to identify several SARS-Cov-2 proteins structures.

Richardson et al. (2020) have used the biomedical knowledge graph method and predicted that Baricitinib –a drug that is often used for the treatment of arthritis– can be used against COVID-19 since this drug inhibits the AP2-associated protein enzyme. As a result, it would be harder for the virus to enter host cells. Hofmarcher et al. (2020) have screened nearly 900 million compounds to estimate their efficacy for the inhibition of 3C and papain-like proteases using a long short-term memory model (Hochreiter and Schmidhuber, 1997). They used important factors, such as toxicity, predicted inhibitory effects, and proximity to known compounds in order to rank them. Finally, they selected 30 000 candidates for further screening.

Regarding the detection of coronavirus, AI methods for automated classification of COVID-19 on computed tomography scans are up-and-coming, as shown by numerous articles (Kundu et al., 2020; Elaziz et al., 2020; Li et al., 2020). Several companies have started developing AI-based apps that work as a COVID-19 health passport, which shows people's vaccination records and COVID-19 disease history and possible exposure to the virus based on interactions with people who might have been positive for the virus (Milliard, 2020).

In the social context of the pandemic, it has become important to control the spread of information online. More than 1 000 fake news were spread during COVID-19 generating a major social problem of misinformation about the disease (Naeem, Bhatti and Khan, 2021). Some intelligent techniques such as transformer-based algorithms, NLP, and supervised learning algorithms, have been implemented for data analysis, information extraction and identification of fake news related with COVID-19 pandemic (Gundapu and Mamidi, 2021). De Magistris et al. (2022) present an automatic fake news detection system based on different techniques from machine learning, deep learning and NLP to check medical news and, in particular, the reliability of publications related to the COVID-19 pandemic, the vaccine and the cure. Similarly, Mookdarsanit and Mookdarsanit (2021) develop a NLP model to identify Thai fake news related to COVID-19. A systematic review of articles indexed in journal citation report on e-health to combat COVID-19 developed by Alonso et al. (2021), provides a guide to deepen the applied work of data science and AI in e-health related to the pandemic. Similar analyses have also been conducted by other authors such as HassanAbady and Ganjali (2021); Monaghesh and Hajizadeh (2020); and Doraiswamy et al. (2020).

5.5. Other topics

The progress of personalized medicine is boosted by the development of *omics* technologies (such as genomics, transcriptomics, proteomics and metabolomics). DS and AI are essential to combine diverse types of omics data, analyzing them, and using the resulting models. Omics technology have the potential of providing a more complete view of biology and disease. Related applications may be found for diagnosis (Ma et al., 2020), prognosis (Poirion et al., 2021), and treatment. In this context, Karczewski and Snyder (2018) describe the utility of combining diverse types of data and the potential applications in human health and disease.

Another field that AI is starting to revolutionize is *drug discovery and design*. Heifetz (2022) describes a number of applications of AI, machine learning, and deep learning in drug design. These new approaches accelerate traditional drug design approaches such as: structure- and ligand-based, augmented and multi-objective de novo drug design, SAR and big data analysis, prediction of binding/activity, and ADMET, among others. This book covers cutting-edge techniques and lists the required software. Yang et al. (2019) explain the basic principles of learning tasks of techniques and describe the state-of-the art of AI-assisted pharmaceutical discovery, covering applications in structure- and ligand-based virtual screening, de novo drug design, physicochemical and pharmacokinetic property prediction, and drug repurposing. Similarly, Jing et al. (2018) discuss applications, limitations, and lines of future research, but focusing on deep learning –including convolutional neural networks, recurrent neural networks, and deep auto-encoder networks.

6. Insights and open challenges

This section offers a brief discussion based on the articles analyzed in the previous sections. We focus on drawing insights from the aforementioned articles, identifying open challenges, and proposing future lines of work.

6.1. Insights from the Literature

According to the results of our search in the Google Scholar and Scopus databases, the number of articles applying data science methods and AI algorithms in e-health has being systematically increasing over the last decade.

The raising interest in these methods and algorithms is a direct consequence of the huge amount of available healthcare data, as well as on the growing demand for new methods and tools that support decision-making in increasingly complex healthcare system. For instance, the growing design and use of smart devices that collect health data favors the use of data analytics and AI techniques in diagnostics and diseases prediction.

A wide range of methodologies have been adopted in e-health already, e.g.: modeling, simulation, statistics, machine learning, data visualization, etc. In particular, it is relevant to highlight the use of ANN in general healthcare and psychological interventions, as well as the use of machine learning to disease prediction. There are several authors that propose an interactive machine learning approach, in which algorithms interact with human or computational agents to optimize the learning process and improve the results. Given the large amount of unstructured data in the sector (e.g., clinical data and diagnostic information in text), NLP techniques are gaining prominence.

Among the different healthcare areas, those where the use of AI can be more valuable are patient administration, clinical decision support, patient monitoring, and healthcare interventions. The most popular topics studied with AI are related to cancer, depression, Alzheimer disease, heart failure, and diabetes. It is commonly accepted that the emergence of data science/analytics and AI represent a paradigm shift in healthcare. Today,

the biggest challenge is a cultural one: it is needed that healthcare staff adopt and employ these techniques regularly. This can be achieved by reducing the gap between AI experts and healthcare staff. Among potential strategies, designing multidisciplinary curricula is the most promising one. Another challenge, regarding the practical implementation of these methods and algorithms, is the lack of a culture on data sharing among patients, hospitals, academia, and industry. Algorithms' transparency and interpretability constitute other barriers to the practical implementation of these approaches.

During the last years, the COVID-19 pandemic has evidenced the key role of data science/analytics methods and IA algorithms in e-health. A large number of authors have made related contributions aiming to identify and visualize the molecular structures of the SARS CoV-2 proteins, evaluate different existing drugs, design new medicines that may help to control the disease, and develop mechanism to detect coronavirus. Indeed, there are many big areas of applications where these analytical and computational tools are extensively being used, among others: patient care (e.g., in diagnosis, prescription, personalized medications and care, etc), public health (e.g., screening, epidemics, forecasting demand for emergency department services, etc), research and development (e.g., drug discovery, gene analysis and editing, etc), healthcare management (e.g., medicine supply chain network design, and treatment/surgery scheduling, etc).

6.2. Open challenges and future research lines

The growing volume of data in healthcare and the increasing complexity of decision-making processes is due, at least in part, to the variability in the evolution of diseases and their interaction with individuals, and gives rise to several challenges. The analysis performed in previous sections shows how data science/analytics and AI methods can efficiently support decision-making and disease diagnosis in different healthcare fields and their potential in the training of medical processes, disease control, and other many areas. According to the OECD (2020), the main challenges facing the integration of intelligent data-driven systems in the healthcare sector are associated with the heterogeneity of health and medical data. In this sector, data are not standardized, varying both between individuals and between populations and subfields. This can create cultural, racial, or geographic biases when transferring the applicability of models to patients or populations with different characteristics from the training data. There is also a risk that the quality and quantity of the data may be far from optimal, thus generating confusion between the noise and the real data during model training and preventing generalized models' development. Besides, much of the data provided in this sector is influenced by the human factor, as practitioners are typically responsible for providing the information. This results in errors, mistakes, and biases in the data, which affects the quality of the learning models.

One of the biggest barriers to the integration of data-driven systems in healthcare is the confidentiality and security of patient data. Information security laws around the world seek after a hazard- and process-oriented approach to guarantee the privacy, astuteness and accessibility of information and the strength of frameworks. This requires

an intermittent handle to audit the adequacy of the security measures and their ceaseless enhancement. Information assurance is not a one-off action, but a task that must be inserted into all exercises relating to the management of health information systems. Similarly, information security could be an assignment and obligation of everybody involved in information handling, and ought to not be allotted solely to an information security officer or information administration division (Organization et al., 2021).

The safe and successful exchange of information must take into account protection issues. The implementation of the EU's Common Information Assurance Control (GDPR) has contributed to hindering the exchange of wellbeing information with analysts outside the EU/EEA. The Common Information Security Control (GDPR) addresses individual information security within the European Union (EU) and European Financial Area (EEA) and the universal exchange of information with areas outside the locality. Additionally, the use of GDPR has presented obstacles to this worldwide exchange of information with outside the EU/EEA, posing problems for academic analysts, healthcare professionals and others within the open division. The European Patients' Group has published data for patients on their rights to information and how exemptions to consent for research purposes (with specialized and authorized shields) should be monitored. The European Commission disseminates a master guideline for EU analysts on morality and information assurance, including universal information exchange. The Chamber of Universal Organizations of Therapeutic Sciences (CIOMS), in collaboration with WHO (CIOMS 2016), has created moral standards for health-related research consent (counting consent for unspecified future use) for the collection, capacity and use of organic tissue and related information. According to the GDPR, consent must be given unreservedly, in particular, educated and unambiguous. All this information is available at the report *International sharing of personal health data for research*, published by All European Academies (ALLEA), the European Academies' Science Advisory Council (EASAC), and the Federation of European Academies of Medicine (FEAM).

According to Blume (2015), the free development of information from the EEA is allowed in case there is a "suitability" choice for the beneficiary. The necessities for a nation to comply with the breadth guidelines are strict (Anghel and Drachenberg, 2019) and depend on whether solid safety standards are as of now connected inside that nation. Up to the present, he European Commission has recognized that some countries have satisfactory security (e.g. Andorra, Argentina, Canada (related to trade associations); Faroe Islands, Guernsey, Israel, Isle of Man, Japan, Shirt, Modern Zealand, Switzerland, and Uruguay). Therefore, there is no broad option for major research-intensive countries, such as China, Australia, USA and South Africa, and it is exceptionally unlikely to occur in countries that lack a legal framework to ensure protection, such as Australia or China. However, it is well known that intelligent methodologies must ensure algorithms transparency, robustness, and security. Risk management approaches (Sahoo et al., 2014) and best practices in applying the methodologies could ensure the protection and responsible use of data. Essential to this is political engagement –both nationally and internationally– in the healthcare sector through policy and regulatory re-

forms that compensate for data protection and the use of data in AI to deliver value to patients and society.

Another barrier is related to the rejection of the integration of AI tools into processes by medical practitioners. Healthcare professionals feel that digital tools could interfere with their patient care decisions. Understanding how they work is essential to influence their perspective on the use of smart technologies. Wang, Kung and Byrd (2018b) states that a data-driven healthcare organization needs data-driven environments that are well understood, reliable, accessible, and secure. Some strategies for the successful integration of smart technologies start from the foundation of strategic planning aimed at establishing a data-driven culture with a robust protocol that enables effective data utilization at all times. All healthcare personnel must acquire and foster a culture of sharing information horizontally –not only among themselves but also with providers and users, developers, and analysts. Likewise, this requires staff trained and skilled in intelligent data-driven technologies, who know the value of data in intelligent decision-making processes. Given the constant evolution of technologies, cloud computing becomes another requirement to meet the challenges of storage, processing, model development, and analysis of results.

There are several research lines that stem from the lack of multidisciplinary integration strategies between data science/analytics/AI scientists and healthcare experts, as well as the development and implementation of data-driven intelligent systems in the sector. The use of AI fosters the development of personalized systems for the treatment and prescription of patients. The development of personalized care systems allows to automate prescriptions and assist patients that require continuous monitoring. Rather than replacing doctors, this helps to enhance their efforts to care for their patients adequately. Other more generalized applications have come to light recently with the COVID-19 pandemic. Many of these show the potential of these models and tools in predicting and controlling virus dynamics and helping to understand and address the consequences in the healthcare sector and society. These applications can transcend to other health processes, such as predicting disease evolution, the scheduling of healthcare resources in different contexts, the study of long-term social impacts, etc.

7. Conclusions

This work has discussed the role of data science/analytics methods and AI algorithms in e-health, pointing out the emerging research topics, reviewing the existing literature, presenting some of the most popular methods and their applications to the healthcare industry, and highlighting the main challenges that need to be yet addressed in order to boost the use of data-intensive methods and algorithms in healthcare, and make the most out of them.

With the increasing use of mobile devices and sensors, large volumes of data can be collected now in real-time. Likewise, digitization processes are increasing the size of medical databases and the possibilities for searching and processing data in them, includ-

ing enhanced data visualization services. Classification and clustering algorithms allow for more intelligent use of both preventive and reactive treatments. Regression models can also be employed to predict values related to illness trajectories. In the e-health context, one should consider interactive machine learning methods, which require human expertise to make efficient decisions when dealing with complex medical treatments. Apart from guaranteeing the quality of data and solving the heterogeneity-of-sources issue, other aspects that have to be considered when dealing with e-health data are related to the legal and privacy dimensions.

Artificial intelligence algorithms have shown to be effective in enhancing the quality of healthcare services (e.g., to speed up diagnostic processes, to personalize medications, to support surgery operations, etc.), increasing the efficiency of public health systems (e.g., by predicting the evolution of epidemics, by forecasting demand for hospital services, etc.), accelerating drug discovery, and supporting healthcare logistics and performance, among many other applications.

Regarding open challenges, these are probably more related to the need for a cultural change than for a technical evolution. In other words, the acceptance of a data- and algorithm-driven culture is needed instead of the traditional one, which is mainly based on the human subjective opinion. Hence, medical experts and healthcare managers need to get used to working hand in hand with data scientists, who will develop models, analyze data, and support the former when making complex decisions involving many variables. Likewise, the heterogeneity of data sources and the privacy and legal aspects associated with health data are relevant barriers that require a considerable effort to be reduced. All in all, the technology and the algorithms are quite advanced already to make better and more extensive use of them if those barriers are eliminated. Therefore, the next decade is expected to provide us with many novel applications of data science/analytics methods and AI algorithms in the healthcare industry.

Acknowledgments

This work has been partially supported by the Divina Pastora Seguros company.

References

Abdalkareem, Z. A., Amir, A., Al-Betar, M. A., Ekhan, P., and Hammouri, A. I. (2021). Healthcare scheduling in optimization context: A review. *Health and Technology*, 11:445–469.

Abulaish, M., Parwez, M. A. and Jahiruddin (2019). Disease: A biomedical text analytics system for disease symptom extraction and characterization. *Journal of Biomedical Informatics*, 100:103324.

Ahsan, M. and Bartlema, J. (2004). Monitoring healthcare performance by analytic hierarchy process: A developing-country perspective. *International Transactions in Operational Research*, 11(4):465–478.

Al-Garadi, M. A., Mohamed, A., Al-Ali, A. K., Du, X., Ali, I., and Guizani, M. (2020). A survey of machine and deep learning methods for Internet of things (IoT) security. *IEEE Communications Surveys & Tutorials*, 22(3):1646–1685.

Al Idrus, A. (2019). Almirall taps Iktos for AI drug discovery work. Fierce Biotech. https://www.fiercebiotech.com/biotech/almirall-taps-iktos-for-ai-drug-discovery-work#:~:text=Almirall%20is%20enlisting%20artificial%20intelligence.

Aldape-Pérez, M., Alarcón-Paredes, A., Yáñez-Márquez, C., López-Yáñez, I., and Camacho-Nieto, O. (2018). An associative memory approach to healthcare monitoring and decision making. *Sensors*, 18(8):2690.

Alonso, S. G., Marques, G., Barrachina, I., Garcia-Zapirain, B., Arambarri, J., Salvador, J. C., and de la Torre Díez, I. (2021). Telemedicine and e-health research solutions in literature for combatting covid-19: a systematic review. *Health and Technology*, 11(2):257–266.

Amato, F., Cozzolino, G., Mazzeo, A., and Romano, S. (2018). Intelligent medical record management: a diagnosis support system. *International Journal of High Performance Computing and Networking*, 12(4):391–399.

Anghel, S. and Drachenberg, R. (2019). Eprs— european parliamentary research service.

Bai, M., Storer, R. H., and Tonkay, G. L. (2017). A sample gradient-based algorithm for a multiple-OR and PACU surgery scheduling problem. *IISE Transactions*, 49(4):367–380.

Bao, H., Sun, X., Zhang, Y., Pang, B., Li, H., Zhou, L., Wu, F., Cao, D., Wang, J., Turic, B., and Wang, L. (2020). The artificial intelligence-assisted cytology diagnostic system in large-scale cervical cancer screening: A population-based cohort study of 0.7 million women. *Cancer Medicine*, 9(18):6896–6906.

Barbieri, C., Cattinelli, I., Neri, L., Mari, F., Ramos, R., Brancaccio, D., Canaud, B., and Stuard, S. (2019). Development of an artificial intelligence model to guide the management of blood pressure, fluid volume, and dialysis dose in end-stage kidney disease patients: Proof of concept and first clinical assessment. *Kidney Diseases*, 5(1):28–33.

Barrett, M., Boyne, J., Brandts, J., Brunner-La Rocca, H.-P., De Maesschalck, L., De Wit, K., Dixon, L., Eurlings, C., Fitzsimons, D., Golubnitschaja, O., Hageman, A., Heemskerk, F., Hintzen, A., Helms, T. M., Hill, L., Hoedemakers, T., Marx, N., McDonald, K., Mertens, M., Müller-Wieland, D., Palant, A., Piesk, J., Pomazanskyi, A., Ramaekers, J., Ruff, P., Schütt, K., Shekhawat, Y., Ski, C. F., Thompson, D. R., Tsirkin, A., Mierden, K. v. d., Watson, C., and Zippel-Schultz, B. (2019). Artificial intelligence supported patient self-care in chronic heart failure: A paradigm shift from reactive to predictive, preventive and personalised care. *EPMA Journal*, 10(4):445–464.

Baud, R., Rassinoux, A.-M., and Scherrer, J.-R. (1992). Natural language processing and semantical representation of medical texts. *Methods of Information in Medicine*, 31(02):117–125.

Beheshtifar, S. and Alimoahmmadi, A. (2015). A multiobjective optimization approach for location-allocation of clinics. *International Transactions in Operational Research*, 22(2):313–328.

Belkhamsa, M., Jarboui, B., and Masmoudi, M. (2018). Two metaheuristics for solving no-wait operating room surgery scheduling problem under various resource constraints. *Computers & Industrial Engineering*, 126:494–506.

Bertsimas, D. and Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044.

Billings, J., Georghiou, T., Blunt, I., and Bardsley, M. (2013). Choosing a model to predict hospital admission: An observational study of new variants of predictive models for case finding. *BMJ Open*, 3(8):e003352.

Birnbaum, D. (2004). Application of data mining techniques to healthcare data. *Infection Control and Hospital Epidemiology*, 25(8):690–695.

Blansit, K., Retson, T., Masutani, E., Bahrami, N., and Hsiao, A. (2019). Deep learning–based prescription of cardiac MRI planes. *Radiology: Artificial Intelligence*, 1(6):e180069.

Blume, P. (2015). Eu adequacy decisions: the proposed new possibilities. *International Data Privacy Law*, 5(1):34.

Bowles, J. (2020). How canadian AI start-up bluedot spotted coronavirus before anyone else had a clue. Diginomica. https://diginomica.com/how-canadian-ai-start-bluedot-spotted-coronavirus-anyone-else-had-clue.

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., and Mohr, D. C. (2011). Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research*, 13(3):e55.

Carrell, D. S., Schoen, R. E., Leffler, D. A., Morris, M., Rose, S., Baer, A., Crockett, S. D., Gourevitch, R. A., Dean, K. M., and Mehrotra, A. (2017). Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *Journal of the American Medical Informatics Association*, 24(5):986–991.

Chakkor, S., Baghouri, M., Cheker, Z., El Oualkadi, A., el Hangouche, J. A., and Laamech, J. (2021). Intelligent network for proactive detection of Covid-19 disease. In *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, pages 472–478. IEEE.

Channa, A., Popescu, N. and Malik, N. Ur R. (2020). Managing covid-19 global pandemic with high-tech consumer wearables: A comprehensive review. In *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 222–228. IEEE.

Chatterjee, A., Gerdes, M. W., and Martinez, S. G. (2020). Statistical explorations and univariate timeseries analysis on COVID-19 datasets to understand the trend of disease spreading and death. *Sensors*, 20(11):3089.

Chen, E., Jiang, J., Su, R., Gao, M., Zhu, S., Zhou, J., and Huo, Y. (2020a). A new smart wristband equipped with an artificial intelligence algorithm to detect atrial fibrillation. *Heart Rhythm*, 17(5):847–853.

Chen, Y., Dong, Y., Zeng, Y., Yang, X., Shen, J., Zheng, L., Jiang, J., Pu, L., and Bao, Q. (2020b). Mapping of diseases from clinical medicine research: a visualization study. *Scientometrics*, 125(1):171–185.

Chukwu, E. and Garg, L. (2020). A systematic review of blockchain in healthcare: Frameworks, prototypes, and implementations. *IEEE Access*, 8:21196–21214.

Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M. W., and Edelson, D. P. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical Care Medicine*, 44(2):368.

Connatty, S. (2019). Capgemini is selected by Bayer as a core strategic partner to transform its IT landscape. Intrado GlobeNewswire. https://www.globenewswire.com/news-release/2019/12/04/1956359/0/en/Capgemini-is-selected-by-Bayer-as-a-core-strategic-partner-to-transform-its-IT-landscape.html.

Coronato, A., Naeem, M., De Pietro, G., and Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964.

Cowie, M. R., Bax, J., Bruining, N., Cleland, J. G., Koehler, F., Malik, M., Pinto, F., van der Velde, E., and Vardas, P. (2016). e-Health: A position statement of the European Society of Cardiology. *European Heart Journal*, 37(1):63.

Crocamo, C., Viviani, M., Bartoli, F., Carrà, G., and Pasi, G. (2020). Detecting binge drinking and alcohol-related risky behaviours from Twitter's users: An exploratory content-and topology-based analysis. *International Journal of Environmental Research and Public Health*, 17(5):1510.

Cui, Z. L., Kadziola, Z., Lipkovich, I., Faries, D. E., Sheffield, K. M., and Carter, G. C. (2021). Predicting optimal treatment regimens for patients with hr+/her2-breast cancer using machine learning based on electronic health records. *Journal of Comparative Effectiveness Research*, 10(9):777–795.

Dang, L. M., Piran, M., Han, D., Min, K., Moon, H., et al. (2019). A survey on Internet of things and cloud computing for healthcare. *Electronics*, 8(7):768.

Dascalu, D., Paraschiv, I. C., Nicula, B., Dascalu, M., Trausan-Matu, S., and Nuta, A. C. (2019). Intelligent platform for the analysis of drug leaflets using nlp techniques. In *2019 18th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, pages 1–6. IEEE.

Datta, S., Li, Y., Ruppert, M. M., Ren, Y., Shickel, B., Ozrazgat-Baslanti, T., Rashidi, P., and Bihorac, A. (2021). Reinforcement learning in surgery. *Surgery*, 170(1):329–332.

Davenport, T. and Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2):94.

De Magistris, G., Russo, S., Roma, P., Starczewski, J. T., and Napoli, C. (2022). An explainable fake news detector based on named entity recognition and stance classification applied to covid-19. *Information*, 13(3):137.

Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772.

Din, I. U., Almogren, A., Guizani, M., and Zuair, M. (2019). A decade of Internet of Things: Analysis in the light of healthcare applications. *IEEE Access*, 7:89967–89979.

Doraiswamy, S., Abraham, A., Mamtani, R., Cheema, S., et al. (2020). Use of telehealth during the covid-19 pandemic: scoping review. *Journal of medical Internet research*, 22(12):e24087.

Downing, N. S., Cloninger, A., Venkatesh, A. K., Hsieh, A., Drye, E. E., Coifman, R. R., and Krumholz, H. M. (2017). Describing the performance of U.S. hospitals by applying big data analytics. *PLOS ONE*, 12(6):1–14.

Dua, S., Acharya, U. R., and Dua, P. (2014). *Machine learning in healthcare informatics*, volume 56. Springer.

Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., and Sahlol, A. T. (2020). New machine learning method for image-based diagnosis of COVID-19. *PLOS ONE*, 15(6):e0235187.

Ellahham, S. and Ellahham, N. (2019). Use of artificial intelligence for improving patient flow and healthcare delivery. *J Comput Sci Syst Biol*, 12(303):2.

Emanuel, E. J. and Wachter, R. M. (2019). Artificial intelligence in health care: Will the value match the hype? *JAMA*, 321(23):2281–2282.

Enticott, J., Johnson, A., and Teede, H. (2021). Learning health systems using data to drive healthcare improvement and impact: A systematic review. *BMC health services research*, 21(1):1–16.

Eysenbach, G. (2001). What is e-health? *Journal of Medical Internet Research*, 3(2):e20.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–37.

Feldman, K., Davis, D., and Chawla, N. V. (2015). Scaling and contextualizing personalized healthcare: A case study of disease prediction algorithm integration. *Journal of Biomedical Informatics*, 57:377–385.

Ferreri, F., Bourla, A., Mouchabac, S., and Karila, L. (2018). e-addictology: an overview of new technologies for assessing and intervening in addictive behaviors. *Frontiers in psychiatry*, page 51.

Fikar, C. and Hirsch, P. (2017). Home health care routing and scheduling: A review. *Computers & Operations Research*, 77:86–95.

Fikar, C., Juan, A. A., Martinez, E., and Hirsch, P. (2016). A discrete-event driven metaheuristic for dynamic home service routing with synchronised trip sharing. *European Journal of Industrial Engineering*, 10(3):323–340.

Friedman, C. and Elhadad, N. (2014). Natural language processing in health care and biomedicine. In *Biomedical Informatics*, pages 255–284. Springer.

Friedman, C. and Hripcsak, G. (1999). Natural language processing and its future in medicine. *Acad Med*, 74(8):890–5.

Galetsi, P. and Katsaliaki, K. (2020). A review of the literature on big data analytics in healthcare. *Journal of the Operational Research Society*, 71(10):1511–1529.

Ganasegeran, K. and Abdulrahman, S. A. (2020). Artificial intelligence applications in tracking health behaviors during disease epidemics. In *Human Behaviour Analysis Using Intelligent Systems*, pages 141–155. Springer.

Ganguly, A. and Nandi, S. (2016). Using statistical forecasting to optimize staff scheduling in healthcare organizations. *Journal of Health Management*, 18(1):172–181.

Gao, X., Jin, Y., Dou, Q., and Heng, P.-A. (2020). Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8440–8446.

Garg, L., McClean, S. I., Barton, M., Meenan, B. J., and Fullerton, K. (2012). Intelligent patient management and resource planning for complex, heterogeneous, and stochastic healthcare systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(6):1332–1345.

Goldstein, B. A., Navar, A. M., and Carter, R. E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *European Heart Journal*, 38(23):1805–1814.

Goodarzian, F., Hosseini-Nasab, H., and Fakhrzad, M. (2020). A multi-objective sustainable medicine supply chain network design using a novel hybrid multi-objective metaheuristic algorithm. *International Journal of Engineering*, 33(10):1986–1995.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18.

Gundapu, S. and Mamidi, R. (2021). Transformer based automatic covid-19 fake news detection system. *arXiv preprint arXiv:2101.00180*.

Guo, L., Xiao, X., Wu, C., Zeng, X., Zhang, Y., Du, J., Bai, S., Xie, J., Zhang, Z., Li, Y., Wang, X., Cheung, O., Sharma, M., Liu, J., and Hu, B. (2020). Real-time automated

diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). *Gastrointestinal Endoscopy*, 91(1):41 – 51.

Gupta, N. and Gupta, B. (2019). Big data interoperability in e-health systems. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 217–222. IEEE.

Haldorai, A., Ramu, A., and Murugan, S. (2019). Biomedical informatics and computation in urban e-health. In *Computing and Communication Systems in Urban Development*, pages 69–89. Springer.

Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician*, 52(2):112–118.

Harper, A., Mustafee, N., and Feeney, M. (2017). A hybrid approach using forecasting and discrete-event simulation for endoscopy services. In *2017 Winter Simulation Conference (WSC)*, pages 1583–1594. IEEE.

Hashimoto, D. A., Rosman, G., Rus, D., and Meireles, O. R. (2018). Artificial intelligence in surgery: Promises and perils. *Annals of Surgery*, 268(1):70.

HassanAbady, S. E. and Ganjali, R. (2021). Medical informatics applications in covid-19 crisis control: Protocol for systematic literature review. *Frontiers in Health Informatics*, 10(1):56.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., and Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25(1):30–36.

Heifetz, A. (2022). *Artificial Intelligence in Drug Design*. Springer.

Hernán, M. A., Hsu, J., and Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *Chance*, 32(1):42–49.

Ho, A. F. W., To, B. Z. Y. S., Koh, J. M., and Cheong, K. H. (2019). Forecasting hospital emergency department patient volume using internet search data. *IEEE Access*, 7:93387–93395.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hofmarcher, M., Mayr, A., Rumetshofer, E., Ruch, P., Renz, P., Schimunek, J., Seidl, P., Vall, A., Widrich, M., Hochreiter, S., and Klambauer, G. (2020). Large-scale ligand-based virtual screening for SARS-Cov-2 inhibitors using a deep neural network. Technical report, SSRN.

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

Hoot, N. R. and Aronsky, D. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2):126–136.

Hoot, N. R., Zhou, C., Jones, I., and Aronsky, D. (2007). Measuring and forecasting emergency department crowding in real time. *Annals of Emergency Medicine*, 49(6):747–755.

Hrinivich, W. T. and Lee, J. (2020). Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning. *Medical physics*, 47(12):6140–6150.

Hu, C., Ju, R., Shen, Y., Zhou, P., and Li, Q. (2016). Clinical decision support for alzheimer's disease based on deep learning and brain network. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE.

Huang, E. and Chang, C.-C. A. (2012). Patient-oriented interactive e-health tools on us hospital web sites. *Health marketing quarterly*, 29(4):329–345.

Huang, E., Chang, C.-c. A., and Khurana, P. (2012a). Users' preferred interactive e-health tools on hospital web sites. *International journal of pharmaceutical and healthcare marketing*.

Huang, T., Lan, L., Fang, X., An, P., Min, J., and Wang, F. (2015). Promises and challenges of big data computing in health sciences. *Big Data Research*, 2(1):2–11.

Huang, Z., Zhou, A., and Zhang, G. (2012b). Non-negative matrix factorization: A short survey on methods and applications. In *International Symposium on Intelligence Computation and Applications*, pages 331–340. Springer.

Hund, M., Sturm, W., Schreck, T., Ullrich, T., Keim, D., Majnaric, L., and Holzinger, A. (2015). Analysis of patient groups and immunization results based on subspace clustering. In *International Conference on Brain Informatics and Health*, pages 358–368. Springer.

Idrees, M. and Sohail, A. (2021). Bio-algorithms for the modeling and simulation of cancer cells and the immune response. *Bio-Algorithms and Med-Systems*, 17(1):55–63.

Impedovo, D., Pirlo, G., and Vessio, G. (2018). Dynamic handwriting analysis for supporting earlier parkinson's disease diagnosis. *Information*, 9(10):247.

Iram, A. and Gill, S. H. (2018). Similarity measuring for clustering patient's reports in telemedicine. In *International Conference on Intelligent Technologies and Applications*, pages 38–49. Springer.

Iroju, O. G. and Olaleke, J. O. (2015). A systematic review of natural language processing in healthcare. *International Journal of Information Technology and Computer Science*, 7(8):44–50.

Javaid, M., Haleem, A., Singh, R. P., and Suman, R. (2021). Dentistry 4.0 technologies applications for dentistry during COVID-19 pandemic. *Sustainable Operations and Computers*, 2:87–96.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., and Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2(4):230–243.

Jing, Y., Bian, Y., Hu, Z., Wang, L., and Xie, X.-Q. S. (2018). Deep learning for drug design: An artificial intelligence paradigm for drug discovery in the big data era. *The AAPS Journal*, 20(3):1–10.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Medicare fraud detection using neural networks. *Journal of Big Data*, 6(1):63.

Jothi, N., Rashid, N., and Husain, W. (2015). Data mining in healthcare–a review. *Procedia Computer Science*, 72:306–313.

Jumper, J., Tunyasuvunakool, K., Kohli, P., Hassabis, D., and the AlphaFold Team (2020). Computational predictions of protein structures associated with COVID-19. Deepmind website. https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19.

Kam, H. J., Sung, J. O., and Park, R. W. (2010). Prediction of daily patient numbers for a regional emergency medical center using time series analysis. *Healthcare Informatics Research*, 16(3):158–165.

Karczewski, K. J. and Snyder, M. P. (2018). Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299–310.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, 15:104–116.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V. A., Wen, C., Zhang, E. D., Zhang, C. L., Li, O., Wang, X., Singer, M. A., Sun, X., Xu, J., Tafreshi, A., Lewis, M. A., Xia, H., and Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.

Khaldi, R., El Afia, A., and Chiheb, R. (2019). Forecasting of weekly patient visits to emergency department: real case study. *Procedia computer science*, 148:532–541.

Kim, D., You, S., So, S., Lee, J., Yook, S., Jang, D. P., Kim, I. Y., Park, E., Cho, K., Cha, W. C., Shin, D. W., Cho, B. H., and Park, H.-K. (2018). A data-driven artificial intelligence model for remote triage in the prehospital environment. *PLOS ONE*, 13(10):e0206006.

Kim, J., Lee, S., Hwang, E., Ryu, K. S., Jeong, H., Lee, J. W., Hwangbo, Y., Choi, K. S., Cha, H. S., et al. (2020). Limitations of deep learning attention mechanisms in clinical research: Empirical case study based on the korean diabetic disease setting. *Journal of Medical Internet Research*, 22(12):e18418.

Koh, H. C. and Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management*, 19(2):65.

Kriegel, H.-P., Borgwardt, K. M., Kröger, P., Pryakhin, A., Schubert, M., and Zimek, A. (2007). Future trends in data mining. *Data Mining and Knowledge Discovery*, 15(1):87–97.

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D., and Ziegler, A. (2014). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal*, 56(4):534–563.

Kshirsagar, R., Hsu, L.-Y., Greenberg, C. H., McClelland, M., Mohan, A., Shende, W., Tilmans, N. P., Guo, M., Chheda, A., Trotter, M., Ray, S., and Alvarado, M.

(2020). Accurate and interpretable machine learning for transparent pricing of health insurance plans. *arXiv preprint arXiv:2009.10990*.

Kundu, S., Elhalawani, H., Gichoya, J. W., and Kahn Jr, C. E. (2020). Radiological Society of North America. How might AI and chest imaging help unravel COVID-19's mysteries? *Radiology: Artificial Intelligence*, 2(3):e200053.

Kwak, G., Ling, L., and Hui, P. (2021). Deep reinforcement learning approaches for global public health strategies for COVID-19 pandemic. *PLoS ONE*, 16(5):e0251550.

Lagos-Ortiz, K., Jácome-Murillo, E., Aguirre-Munizaga, M., and Medina-Moreira, J. (2020). Analysis of the services generated through mobile applications for an accurate diagnosis of epidemiological metrics related to covid-19. In *International Conference on Technologies and Innovation*, pages 151–165. Springer.

Lakshmanaprabu, S., Mohanty, S. N., Krishnamoorthy, S., Uthayakumar, J., Shankar, K., et al. (2019). Online clinical decision support system using optimal deep neural networks. *Applied Soft Computing*, 81:105487.

Lan Hing Ting, K., Voilmy, D., De Roll, Q., Iglesias, A., and Marfil, R. (2021). Field-work and field trials in hospitals: Co-designing a robotic solution to support data collection in geriatric assessment. *Applied Sciences*, 11(7):3046.

Larson, D. and Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, 36(5):700–710.

Lathrop, R. H. (1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Engineering, Design and Selection*, 7(9):1059–1068.

Lepenioti, K., Bousdekis, A., Apostolou, D., and Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, 50:57–70.

Li, K., Fang, Y., Li, W., Pan, C., Qin, P., Zhong, Y., Liu, X., Huang, M., Liao, Y., and Li, S. (2020). CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19). *European Radiology*, 30:4407–4416.

Liang, Y., Yin, S., Tang, Q., Zheng, Z., Elgendi, M., and Chen, Z. (2020). Deep learning algorithm classifies heartbeat events based on electrocardiogram signals. *Frontiers in Physiology*, 11:1255.

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22.

Lisboa, P. J. and Taktak, A. F. (2006). The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks*, 19(4):408–415.

Liu, Y. H. (2017). *Python Machine Learning By Example; Chapter 10: Discovering underlying topics in the newsgroups dataset with clustering and topic modeling*. Packt Publishing Ltd.

Liu, Z., Yao, C., Yu, H., and Wu, T. (2019). Deep reinforcement learning with its application for lung cancer detection in medical internet of things. *Future Generation Computer Systems*, 97:1–9.

Lostumbo, M., Saiz, M., Calvet, L., Juan, A. A., and Lopez, D. (2021). Combining simulation with reliability analysis in supply chain project management under uncertainty: a case study in healthcare. In *Proceedings of the 2021 winter simulation conference*. IEEE.

Lv, J., Deng, S., and Zhang, L. (2020). A review of artificial intelligence applications for antimicrobial resistance. *Biosafety and Health*, 3(1):22–31.

Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., and Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine*, 121:103761.

Makhdoom, I., Abolhasan, M., Lipman, J., Liu, R. P., and Ni, W. (2018). Anatomy of threats to the internet of things. *IEEE Communications Surveys & Tutorials*, 21(2):1636–1675.

Mamdouh, M., Awad, A. I., Hamed, H. F., and Khalaf, A. A. (2020). Outlook on security and privacy in ioht: Key challenges and future vision. In *The International Conference on Artificial Intelligence and Computer Vision*, pages 721–730. Springer.

Martin, C. M., Vogel, C., Grady, D., Zarabzadeh, A., Hederman, L., Kellett, J., Smith, K., and O'Shea, B. (2012). Implementation of complex adaptive chronic care: The Patient Journey Record system (PaJR). *Journal of Evaluation in Clinical Practice*, 18(6):1226–1234.

Martins, L. C., Castaneda, J., Juan, A. A., Barrios, B., Calvet, L., Tondar, A., and Sanchez, J. (2021). Supporting efficient assignment of medical resources in cancer treatments with simulation-optimization. In *Proceedings of the 2021 winter simulation conference*. IEEE.

Martuscelli, C. (2018). AstraZeneca announces China partnerships. MarketWatch. https://www.marketwatch.com/story/astrazeneca-announces-china-partnerships-2018-02-02.

Matheny, M. E., Whicher, D., and Israni, S. T. (2020). Artificial intelligence in health care: A report from the National Academy of Medicine. *JAMA*, 323(6):509–510.

McIntosh, A. M., Stewart, R., John, A., Smith, D. J., Davis, K., Sudlow, C., Corvin, A., Nicodemus, K. K., Kingdon, D., Hassan, L., Hotopf, M., Lawrie, S. M., Russ, T. C., Geddes, J. R., Wolpert, M., Wölbert, E., and Porteous, D. J. (2016). Data science for mental health: A UK perspective on a global challenge. *The Lancet Psychiatry*, 3(10):993–998.

Mesquita, C. T., Oliveira, A., Seixas, F. L., and Paes, A. (2020). Infodemia, fake news and medicine: Science and the quest for truth.

Milliard, M. (2020). Health passports and distancing tools among COVID-19 tech climbing Gartner Hype Cycle. Healthcare IT News. https://www.healthcareitnews.com/news/health-passports-distancing-tools-among-covid-19-tech-climbing-gartner-hype-cycle.

Miori, V. and Russo, D. (2012). Anticipating health hazards through an ontology-based, iot domotic environment. In *2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pages 745–750. IEEE.

Mishra, N. and Silakari, S. (2012). Predictive analytics: A survey, trends, applications, oppurtunities & challenges. *International Journal of Computer Science and Information Technologies*, 3(3):4434–4438.

Monaghesh, E. and Hajizadeh, A. (2020). The role of telehealth during covid-19 outbreak: a systematic review based on current evidence. *BMC public health*, 20(1):1–9.

Mookdarsanit, P. and Mookdarsanit, L. (2021). The covid-19 fake news detection in thai social texts. *Bulletin of Electrical Engineering and Informatics*, 10(2):988–998.

Moreira, M. W., Rodrigues, J. J., Carvalho, F. H., Chilamkurti, N., Al-Muhtadi, J., and Denisov, V. (2019a). Biomedical data analytics in mobile-health environments for high-risk pregnancy outcome prediction. *Journal of Ambient Intelligence and Humanized Computing*, 10(10):4121–4134.

Moreira, M. W., Rodrigues, J. J., Kumar, N., Al-Muhtadi, J., and Korotaev, V. (2018). Nature-inspired algorithm for training multilayer perceptron networks in e-health environments for high-risk pregnancy care. *Journal of Medical Systems*, 42(3):51.

Moreira, M. W., Rodrigues, J. J., Kumar, N., Saleem, K., and Illin, I. V. (2019b). Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. *Information Fusion*, 47:23 – 31.

Moskowitz, D. S. and Young, S. N. (2006). Ecological momentary assessment: What it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience*, 31(1):13.

Murray, E., Hekler, E. B., Andersson, G., Collins, L. M., Doherty, A., Hollis, C., Rivera, D. E., West, R., and Wyatt, J. C. (2016). Evaluating digital health interventions: Key questions and approaches. *American Journal of Preventive Medicine*, 51(5):843–851.

Mutlag, A. A., Abd Ghani, M. K., Arunkumar, N. A., Mohammed, M. A., and Mohd, O. (2019). Enabling technologies for fog computing in healthcare IoT systems. *Future Generation Computer Systems*, 90:62–78.

Naeem, S. B., Bhatti, R., and Khan, A. (2021). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*, 38(2):143–149.

Nandankar, P., Thaker, R., Mughal, S. N., Saidireddy, M., Linda, A., Kostka, J., and Nag, M. A. (2021). An IoT based healthcare data analytics using fog and cloud computing. *Turkish Journal of Physiotherapy and Rehabilitation*, 32:3.

Naslund, J. A., Gonsalves, P. P., Gruebner, O., Pendse, S. R., Smith, S. L., Sharma, A., and Raviola, G. (2019). Digital innovations for global mental health: Opportunities for data science, task sharing, and early intervention. *Current Treatment Options in Psychiatry*, 6(4):337–351.

Neubert, S., Geißler, A., Roddelkopf, T., Stoll, R., Sandmann, K.-H., Neumann, J., and Thurow, K. (2019). Multi-sensor-fusion approach for a data-science-oriented preventive health management system: Concept and development of a decentralized

data collection approach for heterogeneous data sources. *International Journal of Telemedicine and Applications*, 2019(3):1–18.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future: big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375(13):1216.

OECD (2020). Background paper for the G20 AI dialogue, digital economy task force. Technical report.

Oh, H., Rizo, C., Enkin, M., and Jadad, A. (2005). What is eHealth (3): A systematic review of published definitions. *Journal of Medical Internet Research*, 7(1):e1.

Oke, S. (2008). A literature review on artificial intelligence. *International Journal of Information and Management Sciences*, 19(4):535–570.

Oliver, E., Vallés-Pérez, I., Baños, R.-M., Cebolla, A., Botella, C., and Soria-Olivas, E. (2018). Visual data mining with self-organizing maps for "self-monitoring" data analysis. *Sociological Methods & Research*, 47(3):492–506.

O'Neill, J. (2014). *Antimicrobial resistance: Tackling a crisis for the health and wealth of nations*. Review on Antimicrobial Resistance.

Ordu, M., Demir, E., Tofallis, C., and Gunal, M. M. (2021). A novel healthcare resource allocation decision support tool: A forecasting-simulation-optimization approach. *Journal of the operational research society*, 72(3):485–500.

Organization, W. H. et al. (2021). The protection of personal data in health information systems-principles and processes for public health. Technical report, World Health Organization. Regional Office for Europe.

Oztekin, A. (2018). Creating a marketing strategy in healthcare industry: A holistic data analytic approach. *Annals of Operations Research*, 270(1):361–382.

Parfenenko, Y., Verbytska, A., Bychko, D., and Shendryk, V. (2020). Application for medical misinformation detection in online forums. In *2020 International Conference on e-Health and Bioengineering (EHB)*, pages 1–4. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pereira, C. R., Pereira, D. R., Weber, S. A., Hook, C., De Albuquerque, V. H. C., and Papa, J. P. (2019). A survey on computer-assisted parkinson's disease diagnosis. *Artificial intelligence in medicine*, 95:48–63.

Pfeiffer, D. U. and Stevens, K. B. (2015). Spatial and temporal epidemiological analysis in the big data era. *Preventive Veterinary Medicine*, 122(1-2):213–220.

Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S., and Garmire, L. X. (2021). Deepprog: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome medicine*, 13(1):1–15.

Popowich, F. (2005). Using text mining and natural language processing for health care claims processing. *ACM SIGKDD Explorations Newsletter*, 7(1):59–66.

Pore, A., Corsi, D., Marchesini, E., Dall'Alba, D., Casals, A., Farinelli, A., and Fiorini, P. (2021). Safe reinforcement learning using formal verification for tissue retraction

in autonomous robotic-assisted surgery. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4025–4031.

Puliafito, C., Mingozzi, E., Longo, F., Puliafito, A., and Rana, O. (2019). Fog computing for the internet of things: A survey. *ACM Transactions on Internet Technology (TOIT)*, 19(2):1–41.

Pulido, C. M., Ruiz-Eugenio, L., Redondo-Sama, G., and Villarejo-Carballido, B. (2020). A new application of social impact in social media for overcoming fake news in health. *International journal of environmental research and public health*, 17(7):2430.

Quarteroni, A., Lassila, T., Rossi, S., and Ruiz-Baier, R. (2017). Integrated heart-coupling multiscale and multiphysics models for the simulation of the cardiac function. *Computer Methods in Applied Mechanics and Engineering*, 314:345–407.

Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, 2(1):3.

Rais, A. and Viana, A. (2011). Operations research in healthcare: A survey. *International Transactions in Operational Research*, 18(1):1–31.

Rao, A. and Venkatesh, S. (2021). Identification of aphasia using natural language processing. *Journal of University of Shanghai for Science and Technology*, 23:1737–1747.

Ray, E. L. and Reich, N. G. (2018). Prediction of infectious disease epidemics via weighted density ensembles. *PLOS Computational Biology*, 14(2):e1005910.

Razzaque, A. and Hamdan, A. (2020). Artificial intelligence based multinational corporate model for EHR interoperability on an e-health platform. In *Artificial Intelligence for Sustainable Development: Theory, Practice and Future Applications*, pages 71–81. Springer.

Reddy, S., Fox, J., and Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine*, 112(1):22–28.

Reinig, M. (2021). Boehringer & google accelerate drug design with quantum computing. https://www.boehringer-ingelheim.com/press-release/partnering-google-quantum-computing. 03-02-2021.

Richardson, P., Griffin, I., Tucker, C., Smith, D., Oechsle, O., Phelan, A., and Stebbing, J. (2020). Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *Lancet*, 395(10223):e30.

Riise, A., Mannino, C., and Burke, E. K. (2016). Modelling and solving generalised operational surgery scheduling problems. *Computers & Operations Research*, 66:1–11.

Rong, G., Mendez, A., Assi, E. B., Zhao, B., and Sawan, M. (2020). Artificial intelligence in healthcare: Review and prediction case studies. *Engineering*, 6(3):291–301.

Roy, S., Ashaduzzaman, M., Hassan, M., and Chowdhury, A. R. (2018). Blockchain for IoT security and management: Current prospects, challenges and future directions. In *2018 5th International Conference on Networking, Systems and Security (NSysS)*, pages 1–9. IEEE.

Rubrichi, S., Battistotti, A., and Quaglini, S. (2014). Patients' involvement in e-health services quality assessment: a system for the automatic interpretation of sms-based patients' feedback. *Journal of biomedical informatics*, 51:41–48.

Ruiz-Zafra, Á., Benghazi, K., Noguera, M., and Garrido, J. L. (2013). Zappa: An open mobile platform to build cloud-based m-health systems. In *Ambient intelligence-software and applications*, pages 87–94. Springer.

Sahoo, S. S., Jayapandian, C., Garg, G., Kaffashi, F., Chung, S., Bozorgi, A., Chen, C.-H., Loparo, K., Lhatoo, S. D., and Zhang, G.-Q. (2014). Heart beats in the cloud: Distributed analysis of electrophysiological 'Big Data' using cloud computing for epilepsy clinical research. *Journal of the American Medical Informatics Association*, 21(2):263–271.

Saranga, H. and Phani, B. (2009). Determinants of operational efficiencies in the Indian pharmaceutical industry. *International Transactions in Operational Research*, 16(1):109–130.

Senders, J. T., Staples, P. C., Karhade, A. V., Zaki, M. M., Gormley, W. B., Broekman, M. L., Smith, T. R., and Arnaout, O. (2018). Machine learning and neurosurgical outcome prediction: A systematic review. *World Neurosurgery*, 109:476–486.

Serra, D., Romero, P., Lozano, M., García-Fernández, I., Liberos, A., Rodrigo, M., Berruezo, A., Bueno-Orovio, A., and Sebastian, R. (2021). Arrhythmic3D: A fast automata-based tool to simulate and assess arrhythmia risk in 3d ventricular models. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4.

Sethia, D., Gupta, D., Saran, H., Agrawal, R., and Gaur, A. (2016). Mutual authentication protocol for secure nfc based mobile healthcard. *IADIS International Journal on Computer Science & Information Systems*, 11(2).

Shah, J. L., Bhat, H. F., and Khan, A. I. (2021). Integration of Cloud and IoT for smart e-healthcare. In *Healthcare Paradigms in the Internet of Things Ecosystem*, pages 101–136. Elsevier.

Shahid, N., Rappon, T., and Berta, W. (2019). Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLOS ONE*, 14(2):e0212356.

Shailaja, K., Seetharamulu, B., and Jabbar, M. (2018). Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pages 910–914. IEEE.

Shamir, R. R., Dolber, T., Noecker, A. M., Walter, B. L., and McIntyre, C. C. (2015). Machine learning approach to optimizing combined stimulation and medication therapies for parkinson's disease. *Brain stimulation*, 8(6):1025–1032.

Shatte, A. B., Hutchinson, D. M., and Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, 49(9):1426–1448.

Silva, E. A. T., Gomez, I. F. L., Arango, J. F. F., Smith, J. W., Ocampo, S. U., and Hidalgo, J. E. (2018). Evaluation of satisfaction and usability of a clinical decision

support system (cdss) targeted for early obstetric risk assessment and patient follow-up. *E-HEALTH 2018 ICT, Society And Human Beings 2018*, page 3.

Singal, A. G., Mukherjee, A., Elmunzer, B. J., Higgins, P. D., Lok, A. S., Zhu, J., Marrero, J. A., and Waljee, A. K. (2013). Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *The American Journal of Gastroenterology*, 108(11):1723.

Snasel, V., Gajdos, P., Abdulla, H. M. D., and Polovincak, M. (2007). Using matrix decompositions in formal concept analysis. *ISIM*, 252.

Spanakis, G., Weiss, G., Boh, B., Lemmens, L., and Roefs, A. (2017). Machine learning techniques in eating behavior e-coaching. *Personal and Ubiquitous Computing*, 21(4):645–659.

Su, Y.-H., Huang, K., and Hannaford, B. (2021). Multicamera 3d viewpoint adjustment for robotic surgery via deep reinforcement learning. *Journal of Medical Robotics Research*, 6(01n02):2140003.

Sun, P. (2020). Security and privacy protection in cloud computing: Discussions and challenges. *Journal of Network and Computer Applications*, 160:102642.

Susanto, H. (2017). Electronic health system: sensors emerging and intelligent technology approach. In *Smart Sensors Networks*, pages 189–203. Elsevier.

Syed, L., Jabeen, S., Manimala, S., and Elsayed, H. A. (2019). Data science algorithms and techniques for smart healthcare using IoT and big data analytics. In *Smart Techniques for a Smarter Planet*, pages 211–241. Springer.

Tamilselvi, R. and Kalaiselvi, S. (2013). An overview of data mining techniques and applications. *International Journal of Science and Research (IJSR), India Online ISSN*, 2(2):2319–7064.

Taylor, R. A., Pare, J. R., Venkatesh, A. K., Mowafi, H., Melnick, E. R., Fleischman, W., and Hall, M. K. (2016). Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data–driven, machine learning approach. *Academic Emergency Medicine*, 23(3):269–278.

Tejedor, M., Woldaregay, A. Z., and Godtliebsen, F. (2020). Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artificial Intelligence in Medicine*, 104:101836.

Terry, M. (2020). IBM and Pfizer believe machine learning can predict Alzheimer's risk. BioSpace. https://www.biospace.com/article/ibm-and-pfizer-believe-machine-learning-can-predict-alzheimer-s-risk/.

Tomar, D. and Agarwal, S. (2013). A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5):241–266.

Triantafyllidis, A., Velardo, C., Chantler, T., Shah, S. A., Paton, C., Khorshidi, R., Tarassenko, L., Rahimi, K., and SUPPORT-HF Investigators (2015). A personalised mobile-based home monitoring system for heart failure: The SUPPORT-HF Study. *International Journal of Medical Informatics*, 84(10):743–753.

Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., and Collins, G. S. (2019). Predictive analytics in health care: How can we know it works? *Journal of the American Medical Informatics Association*, 26(12):1651–1654.

Van Der Aalst, W. (2016). Data science in action. In *Process Mining*, pages 3–23. Springer.

Vanbrabant, L., Braekers, K., Ramaekers, K., and Van Nieuwenhuyse, I. (2019). Simulation of emergency department operations: A comprehensive review of kpis and operational improvements. *Computers & Industrial Engineering*, 131:356–381.

Vedula, S. S., Ishii, M., and Hager, G. D. (2017). Objective assessment of surgical technical skill and competency in the operating room. *Annual Review of Biomedical Engineering*, 19:301–325.

Virkar, S., Kadam, A., Mallick, S., Tilekar, S., and Raut, N. (2021). An e-health patient management system. In *12th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT 2021*, volume 2021-August, pages 41–46.

Waller, M. A. and Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2):77–84.

Wang, C., Zhu, X., Hong, J. C., and Zheng, D. (2019). Artificial intelligence in radiotherapy treatment planning: Present and future. *Technology in Cancer Research & Treatment*, 18:1533033819873922.

Wang, L. and Alexander, C. A. (2020). Big data analytics in medical engineering and healthcare: Methods, advances and challenges. *Journal of Medical Engineering & Technology*, 44(6):267–283.

Wang, L., Shi, H., and Gan, L. (2018a). Healthcare facility location-allocation optimization for China's developing cities utilizing a multi-objective decision support approach. *Sustainability*, 10(12):4580.

Wang, Y. and Hajli, N. (2017). Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, 70:287–299.

Wang, Y., Kung, L., and Byrd, T. A. (2018b). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126:3–13.

Warmerdam, L., Smit, F., van Straten, A., Riper, H., and Cuijpers, P. (2010). Cost-utility and cost-effectiveness of Internet-based treatment for adults with depressive symptoms: Randomized trial. *Journal of Medical Internet Research*, 12(5):e53.

Watts, J., Khojandi, A., Vasudevan, R., and Ramdhani, R. (2020). Optimizing individualized treatment planning for Parkinson's disease using deep reinforcement learning. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5406–5409. IEEE.

Weltz, J., Volfovsky, A., and Laber, E. B. (2022). Reinforcement learning methods in public health. *Clinical Therapeutics*.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., and Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, 12(4):e0174944.

Widmer, R. J., Collins, N. M., Collins, C. S., West, C. P., Lerman, L. O., and Lerman, A. (2015). Digital health interventions for the prevention of cardiovascular disease: A systematic review and meta-analysis. In *Mayo Clinic Proceedings*, volume 90, pages 469–480. Elsevier.

Wijnhoven, F. (2021). Challenges of adopting human-centered intelligent systems: An organizational learning approach. In *Human Centred Intelligent Systems*, pages 13–25. Springer.

Yang, X., Wang, Y., Byrne, R., Schneider, G., and Yang, S. (2019). Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical Reviews*, 119(18):10520–10594.

Yu, C., Liu, J., Nemati, S., and Yin, G. (2021). Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36.

Yu, K.-H., Beam, A. L., and Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10):719–731.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J. A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., Bikovsky, R., Halpern, Z., Elinav, E., and Segal, E. (2015). Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094.

Zhang, W. and He, X. (2017). An anomaly detection method for medicare fraud detection. In *2017 IEEE International Conference on Big Knowledge (ICBK)*, pages 309–314.

Zhong, L., Luo, S., Wu, L., Xu, L., Yang, J., and Tang, G. (2014). A two-stage approach for surgery scheduling. *Journal of Combinatorial Optimization*, 27(3):545–556.

Zhu, N., Diethe, T., Camplani, M., Tao, L., Burrows, A., Twomey, N., Kaleshi, D., Mirmehdi, M., Flach, P., and Craddock, I. (2015). Bridging e-health and the internet of things: The sphere project. *IEEE Intelligent Systems*, 30(4):39–46.

Zuest, P. (2019). Novartis and Microsoft announce collaboration to transform medicine with artificial intelligence. Novartis. https://www.novartis.com/news/media-releases/novartis-and-microsoft-announce-collaboration-transform-medicine-artificial-intelligence.

SORT 47 (1) January-June 2023, 129-150 DOI: 10.57645/20.8080.02.3

Optimal threshold of data envelopment analysis in bankruptcy prediction

Michaela Staňková[1] and David Hampel[2]

Abstract

Data envelopment analysis is not typically used for bankruptcy prediction. However, this paper shows that a correctly set up a model for this approach can be very useful in that context. A superefficiency model was applied to classify bankrupt and actively manufactured companies in the European Union. To select an appropriate threshold, the Youden index and the distance from the corner were used in addition to the total accuracy. The results indicate that selecting a suitable threshold improves specificity visibly with only a small reduction in the total accuracy. The thresholds of the best models appear to be robust enough for predictions in different time and economic sectors.

MSC: *90C08, 90C90, 90B50, 90B90.*

Keywords: *Bankruptcy prediction, Data envelopment analysis, ROC curve, Threshold optimization, Validation.*

1. Introduction

Evaluating the financial health of companies has been a substantial topic for decades in corporate finance. A company's financial situation is an important guideline not only for the creditors, shareholders and top management of a company in their decision-making but also for the government because the financial distress and bankruptcy of companies (in particular when a larger number of companies go bankrupt in the same period) bring about serious problems such as unemployment. Therefore, there is a constant demand for an ever more accurate and stable tool for forecasting a company's financial situation.

[1] Corresponding author. Department of Statistics and Operation Analysis, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

[2] Department of Statistics and Operation Analysis, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

Received: January 2022
Accepted: March 2023

In the area of financial health assessment, the most frequent topic is the prediction of bankruptcy. Although other situations can be predicted (detection of financial distress or other risks as in Uddin et al. (2020), Petropoulos et al. (2020) or Peláez, Cao and Vilar 2021)), company bankruptcy is a clearly defined situation. Since the second half of the last century, more attention has been given to predicting the financial situation of a company. Many bankruptcy models were developed, which differ both in the method used and the variables used. All bankruptcy models are based on the assumption that companies have some specific symptoms for some time before bankruptcy. These symptoms (i.e., problems) will be reflected in the company's financial statements. Based on these statements, a large number of financial indicators can be defined, making the forecasting of bankruptcy even more difficult.

1.1. Methods used for bankruptcy prediction

The assessment of the financial health of businesses is based on the simple idea of dividing units into two groups: active (healthy) and bankruptcy. There are many methods for dividing companies into two (or more) groups. The earliest known models, such as Beaver (1966) and Altman (1968), were based on multiple discriminant analysis. Later, the logistic regression (logit) model (Ohlson, 1980) and probit model (Zmijewski, 1984) were used in this research area. In addition to these traditional statistical methods, other approaches are also widely applied today. For example, Chen and Du (2009) adopted neural networks to construct a bankruptcy model. Decision trees or the support vector machine method have also been applied; see Klepáč and Hampel (2016) and Li et al. (2018). The application possibilities and especially the predictive abilities of individual methods are still being discussed and researched (see, for example, Klepáč and Hampel (2018) or Staňková and Hampel (2018)). According to Alaka et al. (2018), a total of eight methods can be considered to be suitable for applications in practice. Namely, these are two representatives of statistical approaches (multiple discriminant analysis and logistic regression) and six artificial intelligence tools (artificial neural network, support vector machines, rough sets, case-based reasoning, decision tree and genetic algorithm). The authors conclude that "no single tool is predominantly better than other tools".

For joint stock companies, other options can be used to predict bankruptcy. Campbell, Hilscher and Szilagyi (2011) addresses logit models and includes variables such as excess stock returns and stock volatility. Eisdorfer (2008) used real options techniques, Hillegeist et al. (2004) introduced their own BSM-Prob model based on the Black-Scholes-Merton option-pricing model, and Xu and Zhang (2009) provided an overview of existing approaches with applications to Japanese listed companies. Wu, Gaunt and Gray (2010) presented a new model based on Altman (1968), Ohlson (1980), Zmijewski (1984), Shumway (2001) and Hillegeist et al. (2004). A comprehensive model based on a multiperiod logit model overperforms the original techniques. Attention is given here to the correct selection of variables, where Tian, Yu and Guo (2015) addresses variable selection by the LASSO method and confirms the variables used by Campbell et al.

(2011). Another direction of research is given by Jones (2017) and involved the gradient boosting model, which is capable of using a large number of predictors.

In recent years, efforts have been made to use the data envelopment analysis (DEA) method in financial health assessment. The DEA method typically serves to evaluate the efficiency of decision-making units (DMUs). In this context, DMUs are divided into two groups – efficient (i.e., DMUs that lie on the efficiency frontier) and inefficient (i.e., DMUs that do not lie on the efficiency frontier). However, it is possible to look at efficiency from the other side and focus on finding very inefficient units, which cannot keep up with competition in the longer term and go bankrupt over time. For this reason, the DEA method can be used as not typical tool (e.g., not included in the list of Alaka et al. (2018)) but as a possible tool for predicting bankruptcy.

When employing the DEA method in bankruptcy prediction, the basic option is to calculate the relative efficiency using the DEA model and then use those values in another classification algorithm. For example, Li, Crook and Andreeva (2014) used this approach. Using the radial variable returns to scale model, they calculated the value of the efficiency of Chinese industrial companies and then used those values in the logistic regression model for bankruptcy prediction. A similar procedure can be found, for example, in the studies by Xu and Wang (2009) and Psillaki, Tsolas and Margaritis (2010). Although these studies suggest interesting results, in this paper, we will focus on the possibilities of classification directly through DEA models.

Currently, two different groups of DEA models are developed as a tool for the classification of bankruptcy and active companies. The common idea of both approaches is to estimate the "bankruptcy frontier". One possibility is to use the Azizi and Ajirlu (2010) approach, where the so-called optimistic view of the efficiency frontier changes to a pessimistic view – the original maximization of the objective function is changed to a minimized criterion (i.e., the so-called bounded DEA model). In this case, two frontiers are estimated, which makes it possible to limit the interval at which the production units are located. Another possibility, similar to Janová, Vavřina and Hampel (2012), is to use "standard" DEA models, where input variables are minimized and output variables are maximized with the difference that the variables will be split into inputs and outputs, with the result that the least performing companies heading for bankruptcy appear on the frontier. Active companies should then be within the set of feasible solutions, i.e., not on the "bankruptcy frontier". This approach will be examined in more detail in this work.

Various studies using the DEA method in the field of bankruptcy prediction actually appear; see, for example, Štefko, Horváthová and Mokrišová (2020), Rowland and Krulicky (2020), Chang et al. (2019), Horváthová and Mokrišová (2018), Li, Crook and Andreeva (2017) and Mendelová and Bieliková (2017), but clear application of the selected DEA model is presented there without further investigation or validation.

Janová et al. (2012) used the additive DEA model to predict bankruptcy for agricultural companies. Their models are based on the financial data of the 75 companies obtained from the Amadeus database (54 bankrupt and 21 active companies). This study

shows promising results for using the DEA method for predicting bankruptcy, because overall 75% of companies were correctly classified using this procedure. Staňková and Hampel (2019) also examined the classification capabilities of the additive DEA model for the period of one to three years before bankruptcy. In contrast to the study of Janová et al. (2012), Staňková and Hampel (2019) dealt with a more realistic ratio of active and bankrupt companies in the dataset – 95% active and 5% bankrupt companies. In this case, even for the period of three years before bankruptcy, the DEA models have a total accuracy of over 86%, but at the expense of the error rate of classification of bankrupt companies (error Type II was almost 60%).

Among others, Premachandra, Bhabra and Sueyoshi (2009) focused on the impact of the size constraint on the quality of prediction. They set the ratio of bankrupt/non-bankrupt companies from 0.25 to 1. These changes in settings did not affect the error rate for bankruptcy companies, but they changed the error rate for non-bankrupt companies. At a 1:1 ratio, the overall model error rate was reduced to 14%. Premachandra et al. (2009) found that the DEA model outperforms the logit model in evaluating bankruptcy out-of-sample based on total accuracy. Furthermore, the DEA method does not need the large sample size for bankruptcy evaluation that is usually required by such statistical and econometric approaches. This feature was used in financial evaluation, for example, in Staňková and Hampel (2020).

Premachandra, Chen and Watson (2011) focused on finding a possible proper discriminating or assessing function (based on efficient and inefficient frontiers from the additive superefficiency model) as essential if DEA is used in classification problems such as predicting corporate failure. They started with the idea from logistic regression, where the probability of potential insolvency is calculated and the value of 0.5 is then taken as the classification boundary. Their results show that better results are achieved with lower thresholds (recommendation threshold of 0.1).

It is visible that only a few studies have been conducted to find optimal thresholds in DEA models employed for bankruptcy prediction. Farooq and Qamar (2019) declares the existence of a literature gap about thresholds in general, not only for a case of the DEA method but also for typical data-mining approaches. Several researchers, such as Iparragirre et al. (2022) and Staňková (2022), tried to fill this gap, at least in the case of the logistic regression method. Both mentioned studies used the so-called receiver operating characteristic (ROC) curve to optimize the threshold in the logit model.

Analogically to the case of the logit model, where a probability from the interval $\langle 0; 1\rangle$ that will divide bankrupt and active companies is sought, it is possible to set the threshold value for a particular DEA model. The typical output of a standard DEA models is an efficiency score assigned to each unit, which is compared to a frontier. Units lying on a given frontier are considered efficient, or – in the bankruptcy context – active. For such units, the score is typically equal to one. This principle can also be used in the construction of the "bankruptcy frontier", where companies headed for bankruptcy lie on the frontier (i.e., they have the score equal to one). A score of one can therefore be considered as a threshold where units with a score equal to or above this threshold

will be classified as bankrupt (or inefficient). In the case of bankruptcy assessments, the question is whether such an approach is too strict and whether a different threshold setting would lead to a better bankruptcy prediction success rate. In the case of common DEA models, this threshold is taken from the interval $\langle 0; 1 \rangle$. If the superefficiency model is used, it is also possible to consider values higher than 1 as a potential threshold.

1.2. Motivation and contribution

To date, the DEA method as a tool for constructing the bankruptcy frontier for the purpose of classifying bankrupt companies has rarely been addressed. In the abovementioned publications, attention is typically paid to only one model, usually without further justification of the choice of a specific model. In contrast to these studies, in this paper, we will focus on different model settings to find the most suitable model settings for bankruptcy prediction. In addition, in this study, we will also address the issue of imbalance in the number of active and bankrupt companies in the dataset. In all sectors of the economy, there is naturally an imbalance between the ratio of active and bankrupt companies. Models built on datasets reflecting the real distribution of companies on the market then tend to prefer correct classification in the majority group of active companies, which of course makes them more difficult for real applications. Due to this aspect, this article presents a comprehensive view of the investigated issue.

The main aim of this article is to evaluate and validate the optimal setting of superefficiency DEA models with an optimized threshold for bankruptcy prediction. For these purposes, DEA models are estimated with different settings regarding the measurement method, returns to scale, and orientations. Since we assume that the usual approach of the DEA method, where the value 1 is used as a classification threshold, will not be suitable due to the imbalance of the dataset, attention will also be paid to the identification of a threshold that would allow more balancing of the error rate in both groups of companies. When searching for a threshold, various criteria will be used (especially criteria derived from ROC curves). Different criteria will also be used during the actual evaluation of the classification capabilities of the proposed models for up to three years in advance. The proposed procedure will also be verified on other datasets, and the results of the DEA method will be compared with the competitive statistical method of logistic regression.

The structure of this paper is as follows: Section 2 describes the datasets, variables, models and procedures used. The results are then presented, in Section 3 and the best models are validated and compared with the results of a competing logit model. Finally, the empirical results are discussed, and brief conclusions are provided.

2. Materials and methods

Financial (annual accounting) data on engineering companies (NACE Code 28 – manufacture of machinery and equipment) from 2011 to 2013 were collected from the Orbis

database. To achieve a more homogeneous dataset, only small- and medium-sized companies were included. To obtain an adequate number of bankrupt companies, it was necessary to include companies from across the European Union. The dataset includes 953 companies – 902 active and another 51 companies that in 2014 changed their status to bankruptcy. This dataset (including selected variables) has already been used in the article of Staňková and Hampel (2018), where a suitable setting of standard methods was sought. The use of this dataset will therefore allow a direct comparison with a competing method for bankruptcy prediction.

In their previous research, Staňková and Hampel (2018) identified a group of 19 financial indicators that are suitable for predicting the bankruptcy of engineering companies. They verified this group of variables using three different methods, not including all variables in the models, but letting the method perform the elimination. However, the DEA method itself (unlike, for example, logit or decision tree methods) does not include a mechanism for variable elimination. The involvement of a large number of variables causes several problems in DEA models, such as the instability of the bankruptcy frontier and the dimensionality. For this reason, not all 19 recommended variables will be used in this article. With regard to financial theory, nine characteristics representing the four basic groups of financial ratios (i.e., solvency ratios, profitability ratios, liquidity ratios and turnover ratios) were chosen. Table 1 shows the variables used in the analysis.

Table 1. *Overview of the financial variables used, including their formulas.*

Type	Financial indicator	Formula
Input	Current ratio	$\frac{\text{Current assets}}{\text{Current liabilities}}$
Input	Cash flow liquidity	$\frac{\text{Cash flow}}{\text{Current liabilities}}$
Input	Net working capital (mil. EUR)	Current assets – Current liabilities
Input	Return on assets (%)	$\frac{\text{P/L for period (net income)}}{\text{Total assets}}$
Input	EBIT Margin (%)	$\frac{\text{EBIT}}{\text{Operating revenue (turnover)}}$
Input	Stock turnover	$\frac{\text{Operating revenue (turnover)}}{\text{Stock}}$
Input	Interest cover	$\frac{\text{EBIT}}{\text{Interest paid}}$
Output	Credit period (days)	$\frac{\text{Creditors}}{\text{Operating revenue (turnover)}}$
Output	Debt ratio (%)	$\frac{\text{Noncurrent and current liabilities}}{\text{Total assets}}$

There is a certain risk in the DEA method and in working with ratios. Emrouznejad and Amin (2009) stated that one of the main assumptions related to the typical efficiency frontier in the standard DEA model is the assumption of convexity. When using ratios, it is problematic not to violate this assumption. Despite a certain risk of possible devaluation of the results, however, ratios will be used, because financial ratios are typical

for this type of analysis. Other assumptions regarding the production possibility set of options according to Cooper, Seiford and Tone (2007) can be considered to be fulfilled.

In general, it can be assumed that bankrupt companies have a problem in keeping up with the competition. Their products (services) are more difficult to sell, and therefore, companies have a sales problem. This fact is also negatively reflected in the company's financial statements. Production companies can then accumulate stocks; their turnover is reduced, and so on. Given this fact (and because the output variables are those that are maximized in the DEA model), two financial ratios were selected as output variables: the debt ratio and credit period. Bankrupt companies can be expected to have a higher level of indebtedness (more precisely total debt, especially liabilities), and as a result, the debt ratio will increase. The bankruptcy of the company will also negatively impact the operating cash flow, and companies will not be more likely to repay their own liabilities, thus the credit period will be prolonged.

The remaining seven financial indicators represent input variables. This group of variables contains representatives from all recommended groups, i.e., profitability ratios, liquidity ratios, solvency ratios and turnover ratios. In contrast, it can be expected that the value of these variables in the case of bankrupt companies should be lower than in the case of active companies. It can be assumed that healthy companies will be able to sell their stock and will have a higher level of profitability in all respects. Non-bankrupt companies are expected to have more efficiently adapted internal processes and to be sufficiently liquid and able to pay their obligations.

All of the described procedures are performed in MATLAB R2020b and DEA Solver-Pro version 15.

2.1. Employed bankruptcy prediction DEA models

Due to the nature of the analysis, superefficiency models were selected to compare the resulting score for units that appeared on or above the frontier. All models were estimated separately for the period of one, two and three years prior to bankruptcy. Within each period, 22 superefficiency models with different settings were constructed; see Table 2. Both oriented (input and output orientation) models and nonoriented models were considered. Models were estimated under constant and variable returns to scale. Four models were of a radial nature, and the remaining models were slack-based measure models (SBM models). Since bankrupt companies often have negative financial indicators, we decided to take into account the adjustment of standard DEA models into so-called negative data DEA models (ND models). In such models, according to Cooper et al. (2007), financial ratios are adjusted to a required value greater than zero. Furthermore, attention was given to the so-called SBM Max models. The SBM models typically report the worst efficiency scores for inefficient units. This circumstance means that the projected point is the farthest point on the associated frontier. In contrast to standard SBM models, SBM Max models look for the nearest point on the associated bankruptcy frontier. Hence, the efficiency score is maximized here; for details, see Tone (2017).

Table 2. *Overview of DEA models, including their setup.*

Type	Orientation	Returns to scale	Name
Radial (CCR)	Input	Constant	Model 1
Radial (CCR)	Output	Constant	Model 2
Radial (BCC)	Input	Variable	Model 3
Radial (BCC)	Output	Variable	Model 4
SBM	Non-oriented	Constant	Model 5
SBM	Non-oriented	Variable	Model 6
SBM	Input	Constant	Model 7
SBM	Output	Constant	Model 8
SBM	Input	Variable	Model 9
SBM	Output	Variable	Model 10
SBM Max	Non-oriented	Constant	Model 11
SBM Max	Non-oriented	Variable	Model 12
SBM Max	Input	Constant	Model 13
SBM Max	Output	Constant	Model 14
SBM Max	Input	Variable	Model 15
SBM Max	Output	Variable	Model 16
SBM ND	Non-oriented	Variable	Model 17
SBM ND	Input	Variable	Model 18
SBM ND	Output	Variable	Model 19
SBM ND Max	Non-oriented	Variable	Model 20
SBM ND Max	Input	Variable	Model 21
SBM ND Max	Output	Variable	Model 22

2.2. Characteristics of the model quality

To evaluate the success of the model classification, we follow a number of active/bankrupt companies that are on the frontier and not on the frontier. Based on these characteristics, we can calculate the total accuracy as a percentage of correctly classified subjects for all entities. Instead of the overall misclassification rate of the model, we will focus on the Type I and II errors. A Type I error evaluates the number of active companies that were falsely identified as bankrupt companies to all active companies. A Type II error shows how many bankrupt companies were incorrectly classified as active companies in ratio to all bankrupt companies. More details on these calculations can be found, for example, in Klepáč and Hampel (2018).

Based on the values of Type I and Type II errors, the ROC curve can be constructed. The ROC curve is a useful tool for evaluating classifiers based on their performance. In this context, we will deal with so-called sensitivity, defined as one minus Type I error, and specificity, which equals one minus Type II error. The area under the ROC curve

(AUC) criterion is an alternative single-number measure for evaluating the predictive ability of a model. It was proven in Ling, Huang and Zhang (2003) that the AUC value is a better measure than the total accuracy when evaluating and comparing classifiers. The resulting AUC value is between 0.5 and 1, where higher values indicate a more successful predictive ability for a model.

2.3. Optimal threshold determination

It is possible that some incorrectly classified bankrupt companies could be located near the frontier, and a shift of the bankruptcy frontier as expressed by the threshold value could improve the classification abilities of the DEA model. For this purpose, all theoretically possible thresholds are evaluated (i.e., thresholds from 0 to the maximum value of "bankruptcy score" in the individual model with 0.01 step). To find a suitable threshold, total accuracy maximization and two criteria based on ROC curves were selected.

Similar to Chen and Wu (2016), we use the Youden index, which can be represented as the difference between the probability of a sample predicted as positive when it is truly positive and the probability of the sample predicted as positive when it is not positive. A higher Youden index indicates a better ability to avoid failure in binary classification. Practically, for a particular model, we determine specificity and sensitivity values for all the possible thresholds. The Youden index is then calculated as $J = \max(sensitivity + specificity - 1)$.

Another possibility is to measure the distance from a "perfect" model with zero Type I error as well as Type II error (point [0; 1] on the ROC curve) to the nearest point of the ROC curve of the assessed model. This approach produces so-called distance-to-corner characteristics, which correspond to a suitable threshold.

2.4. Validation of the results

For decisions about the possible systematic behaviour of threshold setting, it is necessary to check the stability of the optimal threshold for particular models. For this purpose, we employ additional datasets coming from different time ranges. These consist of company data from 2013, 2014 and 2015, where some companies became bankrupt in 2016. The first dataset consists of companies from the sector NACE Code 28, i.e., the sector used for establishing optimal thresholds. In addition to threshold validation based on data from different time ranges, we employ two datasets from different sectors: the manufacture of basic metals (NACE Code 24) and the manufacture of fabricated metal products, except for machinery and equipment (NACE Code 25). These two sectors are chosen not only for their comparability with the manufacture of machinery and equipment sector but also for the existence of a sufficient number of bankruptcies with available data. The composition of the validation datasets in particular years is presented in Table 3.

Validation will be performed for DEA bankruptcy models with the best classification capabilities within the original dataset. When optimal thresholds based on NACE Code 28 sector data from 2011–2013 will give reasonably good results not only for the same

Table 3. *Composition of validation datasets.*

	Active companies (in %)			Bankrupt companies (in %)		
Year	NACE 24	NACE 25	NACE 28	NACE 24	NACE 25	NACE 28
2013	3180 (99.07)	1485 (92.47)	3187 (97.05)	30 (0.93)	121 (7.53)	97 (2.95)
2014	2784 (98.93)	1472 (92.99)	3136 (97.42)	30 (1.07)	111 (7.01)	83 (2.58)
2015	2810 (99.01)	1479 (94.44)	3188 (97.82)	28 (0.99)	87 (5.56)	71 (2.18)

sector in different time ranges but also for other sectors, we can have good faith that the optimal thresholds found will be applicable in general.

To correctly evaluate the qualification capabilities of the DEA method, the results of the best models will be compared with the results of logistic regression. For this purpose, the already tested model from Staňková and Hampel (2018) will be used. More precisely, it is a model constructed by means of forward stepwise regression – the starting model in this case contains only a constant. Logit models will be estimated for both the original and validation datasets.

3. Results

As the first part of the evaluation of the prediction capabilities of the model, we apply the common approach, where we used the value of one as the threshold for the classification of active and bankrupt companies. It was found that in such a case, the estimated models (except for Model 3) typically have a very low error rate in the group of active companies (values lower than 1%) but a very high error rate in the group of bankrupt companies (typically approximately 80 to 90%). Such models cannot be considered to be models that can be used in practice. It was also found that one of the models had problems with the superefficiency calculation. For Model 22, we were unable to obtain results in any of the three reporting periods. It can be assumed that this model does not have the appropriate settings due to the problem and data being studied.

In the case of a typical dividing point of 1, it was also found that Model 3 is different from the others. Model 3 showed the smallest error rate in the classification of the minority bankruptcy group (i.e., had the lowest value of Type II error in all three periods). However, it lags behind in terms of overall accuracy. Because of its setting, Model 3 (compared to other models) has a large number of companies on or above the bankruptcy frontier. For example, in the period of three years before bankruptcy, there were 298 companies (38 bankrupt and 260 active units). In all other models, only a few units or tens of units appeared on or above the bankruptcy frontier in this period. These specific features of Model 3 caused a dramatic reduction in the overall accuracy of the model to a value of approximately 72% (in all periods), and thus, in terms of total accuracy, Model 3 was the least suitable model. In addition, in all periods, the AUC values of Model 3 exceed the overall accuracy, which indicates that it is advisable to look for a threshold

other than the value threshold. In the period of one year before bankruptcy, this fact applies not only to Model 3 but also to the eight other models.

Generally, the standard threshold equal to 1 does not reach the maximum value of the total accuracy. When choosing a model with a threshold selected to maximize the total accuracy, typically more than one appropriate threshold was found within the model. It was also found that when maximizing the overall accuracy, it is advisable very often to use thresholds higher than one. However, threshold values founded by the criterion of maximum accuracy can be described as inappropriate. For this specific threshold value, we obtain a model where the Type I error is very low but at the expense of correct classification of the less frequent companies that went bankrupt during the observed period, which results in a very low specificity value (and thus a high Type II error).

In the case of characteristics derived from ROC curves, typically one point was found relating to the given criterion. It can be stated that for the period of three years before bankruptcy, the thresholds found by the distance to corner and the Youden index are relatively consistent. Only in 8 cases (9 cases in the period of two years before the bankruptcy) did the identified thresholds differ according to these two criteria. The thresholds found typically differed by only a few hundredths, but for Models 3 and 14, the difference was 0.59 and 0.89 points in the one-year period before bankruptcy.

3.1. The results of the best models

To select the best models, the AUC values were first monitored, while the ranking of the models according to the AUC values in individual years was averaged to create the resulting average ranking of the models for the entire monitored period. Type I and II errors were monitored as a second criterion. From the group of radial models (i.e., Models 1 to 4), Model 4 was selected as the best model. Radial models in terms of AUC values had the most fluctuating results. According to the average values, Model 4 was the best, but when changing the orientation (i.e., changing to Model 3), according to the average AUC values, we obtain one of the worst models (18th in the ranking). From the group of "basic" SBM models, i.e., Models 5 to 10, Model 6 was selected as the best. In general, however, these SBM models achieved very good results (in the average ranking according to AUC, it was the 3rd to 9th position). The other SBM models fared worse in terms of average AUC values. Of the group of SBM MAX models, i.e., Models 11 to 16, Model 12 performed best. From the group of models with special treatment for negative data, i.e., Models 17 to 21, Model 17 can be identified as the best.

Several links can be found among the selected best models from each group. It was found that among the best models are models with variable returns to scale, and in three out of four cases, it is a nonoriented model. Furthermore, these models show that the value of the optimal threshold decreases with the onset of bankruptcy. It can be assumed that as the time of bankruptcy approaches, active companies become more different from bankrupt companies and therefore move away from the bankruptcy frontier, and therefore, the threshold decreases. A detailed view of the thresholds found for the four selected models according to the criterion of maximum accuracy (C1), distance to

corner (C2) and Youden index (C3) for the period of one to three years before bankruptcy can be seen in Table 4. Using the thresholds found, we can make some generalizations. The Youden index and the distance to the corner show the same thresholds, with the exception of certain models for data three years before bankruptcy. Maximizing the overall accuracy gives thresholds that are substantially larger than the other criteria, and the difference decreases with increasing time to bankruptcy. Notable is Model 6, which has optimal thresholds that are mostly very similar across different times to bankruptcy and different criteria.

Table 4. *Thresholds found by the criteria C1 (maximum of total accuracy), C2 (distance to corner) and C3 (Youden index) for selected models. For criterion C1, the same accuracy values were often achieved for different thresholds, so the threshold leading to maximum accuracy is represented in the table as an interval of threshold values. This is a consequence of using an empirical ROC curve which is piecewise constant.*

Period	3 years before bankruptcy			2 years before bankruptcy			1 year before bankruptcy		
Criterion	C1	C2	C3	C1	C2	C3	C1	C2	C3
Model 4	$\langle 3.32, 5.74\rangle$	0.58	0.58	$\langle 0.57, 0.58\rangle$	0.27	0.27	0.14	0.10	0.10
Model 6	$\langle 1.25, 1.27\rangle$	0.07	0.07	0.10	0.01	0.01	0.07	0.01	0.01
Model 12	$\langle 1.25, 1.27\rangle$	0.32	0.49	$\langle 0.67, 0.68\rangle$	0.21	0.23	$\langle 0.24, 0.25\rangle$	0.05	0.05
Model 17	$\langle 1.31, 1.41\rangle$	0.15	0.24	$\langle 1.03, 1.49\rangle$	0.24	0.23	0.09	0.03	0.04

The predictive abilities of the best models from each group described above are depicted in Figure 1. This figure shows four evaluation criteria having a maximization character: the value of area under ROC curve (AUC), total accuracy (ACC), specificity (SPE) and sensitivity (SEN). The results show that the typically used point of one (magenta) as well as the threshold according to the criterion of maximum accuracy (green) lead to models where the emphasis is placed on the correct specification in the group of active companies (i.e., high sensitivity value) at the expense of correct classification in the group bankrupt companies (i.e., low value of specificity). However, if the Youden index and the distance to corner criteria are used, the results in all four evaluation areas are balanced. In cases where these two criteria did not agree on the same dividing point, the Youden index (blue) tended to have a higher sensitivity at the expense of specificity than the distance to corner (black) criterion. For all four selected models (regardless of the specific threshold), it can also be observed that the overall quality of the model decreases as the time since bankruptcy increases.

3.2. Results via validation datasets

For the four DEA models selected above, the classification capabilities of these models were verified on the other three datasets. In those cases where there was no agreement between the distance to corner criterion and the Youden index during the optimization of the threshold, only the distance to corner criterion was uniformly used, which identified

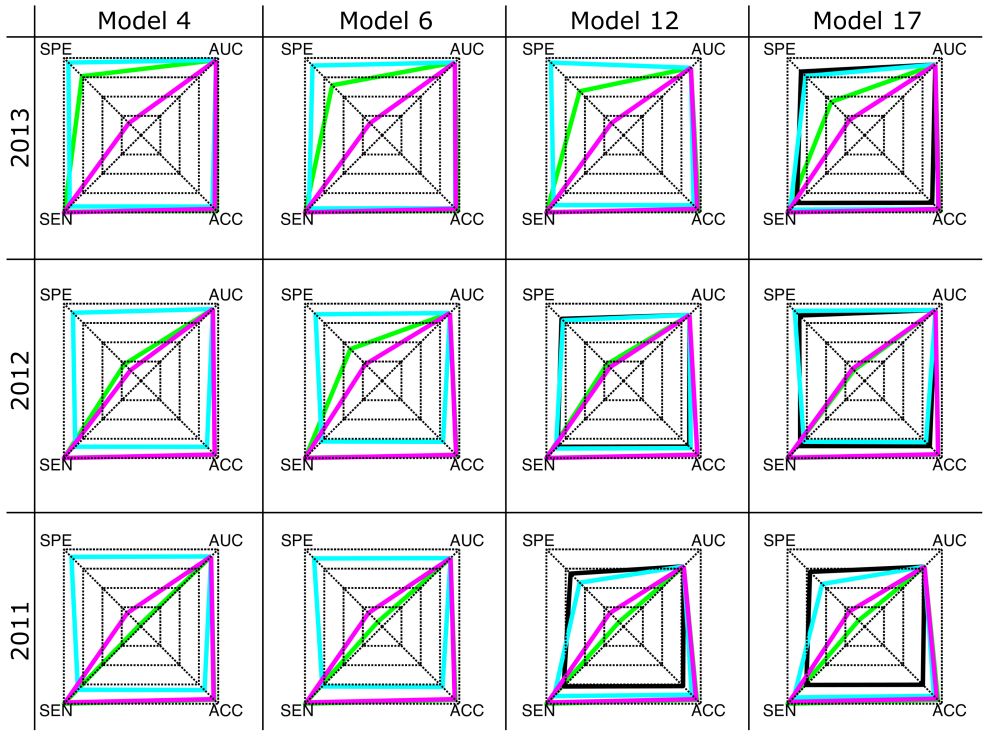

Figure 1. *Results of the area under ROC curve (AUC), total accuracy (ACC), specificity (SPE) and sensitivity (SEN) with a typical threshold equal to one (magenta) and thresholds found by the Youden index (blue), distance to corner (black) and maximum accuracy (green) for selected models. Because the thresholds coincided in some cases, not all characteristics (colours) are visible in the picture.*

thresholds that better balanced both types of errors in the original dataset. The results of total accuracy (ACC), specificity (SPE) and sensitivity (SEN) and AUC for the original dataset and for three validation datasets for one year before bankruptcy (red), two years (green) and three years before bankruptcy (blue) are shown in Figure 2. The last column in Figure 2 then presents the results for the competing logit model. As seen, the logit model based on the original dataset achieved similar results to the DEA models for the period of one year before bankruptcy. However, for other time periods and other sectors, the logit model lags significantly in the specificity values.

If we focus on the evaluation of individual DEA models, then in the case of the first validation set (NACE Code 28), Model 17 failed visibly. The classification abilities of the other models are still very good. Model 6 has the most comparable results to the original dataset. In the case of Model 4 and Model 12, very good results were achieved in the period of one year before bankruptcy, but for periods longer than one year before bankruptcy, a decrease in the specificity values can be seen for both of these models. In the case of the second validation dataset (NACE code 24), the best results were achieved

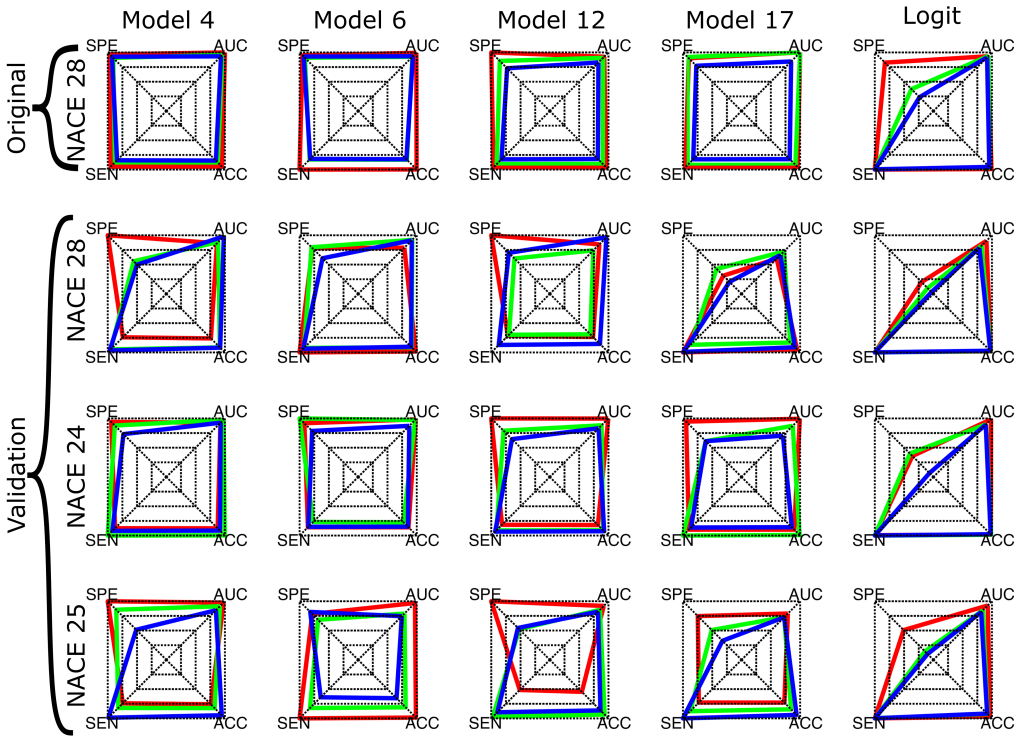

Figure 2. *Validation results for Models 4, 6, 12 and 17 together with the logit model. Original dataset characteristics with optimal threshold for selected models (the first line, NACE Code 28, bankruptcy in 2013), i.e., the area under the ROC curve (AUC), total accuracy (ACC), specificity (SPE) and sensitivity (SEN) are compared to the characteristics of the validation datasets: NACE Code 28, bankruptcy in 2016 (the second line), NACE Code 24, bankruptcy in 2016 (the third line) and NACE Code 25, bankruptcy in 2016 (the fourth line). Depicted are results for one year before bankruptcy (red), two years (green) and three years before bankruptcy (blue).*

with Models 4 and 6 throughout the observed period. Models 12 and 17 were less able to classify bankrupt companies in the period of two or more years before bankruptcy. Even in the case of the third dataset (NACE Code 25), one can see the fluctuation in the results of the models with respect to the remaining time to bankruptcy. If we were to take the classification of bankrupt companies as a priority (i.e., the specificity results), Model 6 was identified as the best.

Given the validation results, it can be stated generally that the thresholds for Model 6 appear to be stable regardless of the sector chosen and the different time periods. Model 4 can be called the second best, and Models 12 and 17 performed the worst during validation. One can assume that in these worst-case models, the optimal threshold will be more influenced by the specific characteristics of bankrupt companies in the sector. It can be said that with increasing time since the bankruptcy of a company, the ideal thresholds are more affected by other influences (industry specificities or directly by the characteristics of bankrupt companies).

4. Discussion

Generally, there is a strictly limited number of research papers dealing with the validation of DEA bankruptcy models, and subsequent threshold optimization is rarely resolved. The DEA method is not yet a broadly accepted method for the area of bankruptcy prediction (Alaka et al., 2018). However, when comparing the empirical results with common logit models (in Figure 2), the proposed DEA models have more potential for practical application. In addition, the DEA method has one advantage over conventional statistical approaches, because it does not require large datasets. This aspect allows the application of threshold-optimized DEA models in relatively small economic sectors or in the case of oligopolies.

The usefulness of threshold optimization enabled by using superefficiency models can be demonstrated by comparison with the results of Janová et al. (2012), Premachandra et al. (2009) and Staňková and Hampel (2019), where additive models are used with a standard threshold corresponding to zero slack values. Threshold optimization using the Youden-like approach of the additive DEA model is elaborated in Štefko et al. (2020). Since the proportion of active companies to bankrupt companies is not balanced in these datasets, not only the total accuracy but also the error rates for both active and bankrupt companies must be accounted for to prevent the loss of error margin classification of the less frequent companies that went bankrupt during the observed period. The characteristics of bankruptcy prediction in the abovementioned sources are summarized in Table 7.

Table 5. *Results of Janová et al. (2012), Premachandra et al. (2009), Staňková and Hampel (2019) and Štefko et al. (2020) and our results for the case one year before bankruptcy. Abbreviation ACC means total accuracy, TIE Type I error and TIIE Type II error.*

Source	ACC	TIE	TIIE
Janová et al. (2012)	0.746	0.003	0.805
Premachandra et al. (2009)	0.686	0.011	0.872
Staňková and Hampel (2019)	0.940	0.029	0.490
Štefko et al. (2020)	0.593	0.446	0.180
The best model via original dataset	0.946	0.051	0.098
The best model via validation NACE 28 dataset	0.834	0.152	0.338
The best model via validation NACE 24 dataset	0.801	0.200	0.143
The best model via validation NACE 25 dataset	0.897	0.090	0.287

It is obvious that the total accuracy of our best model results and the results reached in Staňková and Hampel (2019) are visibly higher than in Janová et al. (2012) and Premachandra et al. (2009). Inter alia, this finding can be given by the different variables used. It can be stated that the identified primary group of ratio indicators in Staňková and Hampel (2018) is suitable not only for the methods of logistic regression, support vector

machines and decision trees but also to serve as a basic set for the DEA method, because both Staňková and Hampel (2019) models and our models achieved good classification results through these variables.

The research of Premachandra et al. (2009) addresses the problem of bankrupt companies' share in the dataset. They show that it is easier for the DEA method to address balanced data files (increase of total accuracy from 75% for the original dataset to 86% for the balanced dataset). We can assume that threshold optimization does not bring a serious advantage in the case of a balanced dataset, but this situation is not real. The strongly unbalanced data truly reflect the situation in today's market, which is populated far more densely by active companies than by those that are on the brink of bankruptcy. Therefore, for such datasets and especially for periods longer than one year before bankruptcy, we do not consider a threshold that is equal to or greater than one to be an optimal setting.

In Štefko et al. (2020), the authors address predicting bankruptcy in the heating industry in Slovakia. The additive DEA model and logit model are employed for this purpose. Threshold optimization based on maximization of the sum of sensitivity and specificity is provided. As in this paper, a set of 9 financial indicators with no strong correlations is used. The dataset consists of 343 companies, of which 50 were bankrupted in 2016. A relatively low total accuracy of 56% is reached, and the type II error is close to our best results, but the type I error is high. In accordance with our approach, the usefulness of threshold optimization is visible.

If we optimize the threshold in our proposed DEA models, we will not achieve the maximum total accuracy of the model, but we will obtain models where both types of errors are more balanced. From this point of view, the models proposed by us are therefore more applicable in practice than, for example, the models by Štefko et al. (2020) and Premachandra et al. (2009). With respect to the identified thresholds and classification capabilities in the original as well as the validation datasets, nonoriented SBM models proved to be the best. In general, better results were achieved by models with the assumption of variable returns to scale; however, in the case of nonoriented SBM models, the change in this setting had no significant effect on the results of the models.

Empirical results showed, among other things, that in the case of criteria derived from ROC curves, it is not advisable to use thresholds higher than 1. There is therefore no need to distinguish between companies that form the bankruptcy frontier. In practice, this means that it is possible to estimate models in their basic form (i.e., without the need to calculate superefficiency scores).

Due to the empirical shape of the estimated ROC curves, the optimal values of thresholds given by the Youden index and distance to corner do not always match exactly. However, these suitable thresholds are usually not very far away from each other. In the event that these two criteria did not provide the same values, the model based on the threshold according to the distance to corner usually balanced both types of errors slightly better. When selecting the criterion, the purpose of the models must be

accounted for. If the user of the model (for example, a bank) is more interested in the correctness of the classification of a minority group of bankrupt companies, we can recommend thresholds given by the distance to corner. A model with a slightly lower overall accuracy but higher specificity will be obtained in this manner.

5. Conclusions

Given the results, it can be stated that threshold optimization can visibly improve the quality of a DEA model's bankruptcy prediction. The selection of a given threshold is individual for each type of DEA model and for the period. However, nonoriented SBM models showed that they generally have relatively low ideal thresholds according to ROC curves in the range of 0.01 to 0.07. Therefore, these models were also marked as the best. These models are the most robust in the sense of the method for optimal threshold determination and the type of returns on scale, and furthermore, these models are stable in the sense of optimal threshold for different periods before bankruptcy. Validation proved that the high quality of nonoriented SBM models' bankruptcy prediction persisted for different sector companies' data. Although we assume that the results obtained will be stable both over time and for different sectors of the economy, it will be useful in the future to check the validity of the results under different circumstances, namely in a different time frame, sector, and country. Future research will also focus on different estimation methods of ROC curves, where we can assume that smooth ROC curves will provide more stable threshold estimates.

Conflict of interest

The authors report that they have no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Author Contributions: Michaela Staňková: investigation, data curation, formal analysis, visualization, writing-original draft; David Hampel: conceptualization, formal analysis, review and editing. All authors have read and agreed to the published version of the manuscript.

References

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O. and Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, 94:164–184.

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609.

Azizi, H. and Ajirlu, S. (2010). Measurement of overall performances of decision-making units using ideal and anti-ideal decision-making units. *Computers & Industrial Engineering*, 59(3):411–418.

Beaver, W. (1966). Financial ratios as predictors of failure. *The Journal of Accounting Research*, 4:71–102.

Campbell, J. Y., Hilscher, J. and Szilagyi, J. (2011). Predicting financial distress and the performance of distressed stocks. *Journal of Investment Management*, 9(2):14–34.

Chang, T. M., Lin, S. J., Chen, F. H. and Hsu, M. F. (2019). Integrated multiple DEA specifications and visualization technique for advanced management analysis and decision. In *2019 20th International Conference on Mobile Data Management (MDM 2019)*, pages 491–496, Hong Kong. IEEE.

Chen, W. S. and Du, Y. K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications*, 36(2):4075–4086.

Chen, Y. and Wu, W. (2016). A prospecting cost-benefit strategy for mineral potential mapping based on ROC curve analysis. *Ore Geology Reviews*, 74:26–38.

Cooper, W. W., Seiford, M. L. and Tone, K. (2007). *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*. Springer Science & Business Media, New York.

Eisdorfer, A. (2008). Empirical evidence of risk shifting in financially distressed firms. *The Journal of Finance*, 63(2):14–34.

Emrouznejad, A. and Amin, G. (2009). DEA models for ratio data: Convexity consideration. *Applied Mathematical Modelling*, 33(1):486–498.

Farooq, U. and Qamar, M. A. J. (2019). Predicting multistage financial distress: Reflections on sampling, feature and model selection criteria. *Journal of Forecasting*, 38(7):632–648.

Hillegeist, S. A., Keating, E. K., Cram, D. P. and Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, 9(1):5–34.

Horváthová, J. and Mokrišová, M. (2018). Risk of bankruptcy, its determinants and models. *Risks*, 6(4):117.

Iparragirre, A., Irantzu, B., Jorge, A. and Inmaculada, A. (2022). Estimation of cut-off points under complex-sampling design data. *SORT-Statistics and Operations Research Transactions*, 46(1):137–158.

Janová, J., Vavřina, J. and Hampel, D. (2012). DEA as a tool for bankruptcy assessment: the agribusiness case study. In *Proceedings of the 30th International Conference Mathematical Methods in Economics 2012*, pages 379–383, Karviná. Silesian University in Opava.

Jones, S. (2017). Corporate bankruptcy prediction: A high dimensional analysis. *Review of Accounting Studies*, 22(3):1366–1422.

Klepáč, V. and Hampel, D. (2016). Prediction of bankruptcy with SVM classifiers among retail business companies in eu. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 64(2):627–634.

Klepáč, V. and Hampel, D. (2018). Predicting bankruptcy of manufacturing companies in EU. *E&M Ekonomie a Management*, 21(1):159–174.

Li, H., Hong, L., Mo, Y., Zhu, B. and Chang, P. (2018). Restructuring performance prediction with a rebalanced and clustered support vector machine. *Journal of Forecasting*, 37(4):437–456.

Li, Z., Crook, J. and Andreeva, G. (2014). Chinese companies distress prediction: an application of data envelopment analysis. *Journal of the Operational Research Societ*, 65(3):466–479.

Li, Z., Crook, J. and Andreeva, G. (2017). Dynamic prediction of financial distress using malmquist DEA. *Expert Syst Appl*, 80:94–106.

Ling, C. X., Huang, J. and Zhang, H. (2003). AUC: A better measure than accuracy in comparing learning algorithms. In Xiang, Y. and Chaib-draa, B., editors, *Advances in Artificial Intelligence. Canadian AI 2003. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, volume 2671. Springer, Berlin, Heidelberg.

Mendelová, V. and Bieliková, T. (2017). Diagnosing of the corporate financial health using DEA: an application to companies in the Slovak republic. *Politicka Ekonomie*, 65(1):26–44.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131.

Peláez, R., Cao, R. and Vilar, J. M. (2021). Nonparametric estimation of the probability of default with double smoothing. *SORT-Statistics and Operations Research Transactions*, 45(2):93–120.

Petropoulos, A., Siakoulis, V., Stavroulakis, E. and Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting*, 36(3):1092–1113.

Premachandra, I. M., Bhabra, G. S. and Sueyoshi, T. (2009). DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique. *European Journal of Operational Research*, 193(2):412–424.

Premachandra, I. M., Chen, Y. and Watson, J. (2011). DEA as a tool for bankruptcy assessment: A case of bankruptcy assessment. *Omega*, 39(6):620–626.

Psillaki, M., Tsolas, I. E. and Margaritis, D. (2010). Evaluation of credit risk based on firm performance. *European Journal of Operational Research*, 201(3):873–881.

Rowland, Z. and Krulicky, T. (2020). Using DEA as a useful tool for bankrupty assessment in Romanian's enterprises. In *7th International Conference on Education and Social Sciences (INTCESS 2020)*, pages 1213–1217, Istanbul. International Organization Center of Academic Research.

SAITECH Inc. (2020). DEA SolverPRO version: 15f.

Shumway, T. (2001). Forecasting bankruptcy more accurately: a simple hazard model. *Journal of Business*, 74(1):101–124.

Staňková, M. (2022). Threshold moving approach with logit models for bankruptcy prediction. *Computational Economics*.

Staňková, M. and Hampel, D. (2018). Bankruptcy prediction of engineering companies in the EU using classification methods. *Acta Universitatis agriculturae et silviculturae Mendelianae Brunensis*, 66(5):1347–1356.

Staňková, M. and Hampel, D. (2019). Bankruptcy prediction based on data envelopment analysis. In *Mathematical Methods in Economics 2019: Conference Proceedings*, pages 31–36, České Budějovice. University of South Bohemia in České Budějovice.

Staňková, M. and Hampel, D. (2020). Efficiency assessment of the UK travel agency companies - data envelopment analysis approach. In *Mathematical Methods in Economics 2020: Conference Proceedings*, pages 550–556, Brno. Mendel University in Brno.

The MathWorks Inc. (2020). MATLAB version: 9.13.0 (r2020b), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com

Tian, S., Yu, Y. and Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52:89–100.

Tone, K. (2017). *Advances in DEA Theory and Applications: With Extensions to Forecasting Models*. John Wiley and Sons, Hoboken.

Uddin, M. S., Chi, G. T., Habib, T. and Zhou, Y. (2020). An alternative statistical framework for credit default prediction. *Journal of Risk Model Validation*, 14(2):65–101.

Štefko, R., Horváthová, J. and Mokrišová, M. (2020). Bankruptcy prediction with the use of data envelopment analysis: An empirical study of Slovak businesses. *Journal of Risk and Financial Management*, 13(9):212.

Wu, Y., Gaunt, C. and Gray, S. (2010). A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting & Economics*, 6(1):34–45.

Xu, M. and Zhang, C. (2009). Bankruptcy prediction: The case of japanese listed companies. *Review of Accounting Studies*, 14(4):534–558.

Xu, X. and Wang, Y. (2009). Financial failure prediction using efficiency as a predictor. *Expert Systems with Applications*, 36(1):366–373.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 22:59–86.

Appendices

Table 6. *Median and average values for the used variables separately for active and bankrupt companies.*

	Median/average values for					
	Active companies			Bankrupt companies		
Variable	2011	2012	2013	2011	2012	2013
Current ratio	1.09/1.21	1.01/1.11	0.75/1.04	1.52/1.88	1.54/1.96	1.57/1.93
Cash flow liquidity	0.03/-0.03	-0.03/-0.13	-0.31/-1.59	0.14/0.24	0.15/0.22	0.15/0.23
Net working capital	0.15/0.34	0.01/-0.34	-0.43/-1.03	15.03/42.57	15.76/49.02	16.19/50.80
Return on assets	-0.13/-5.08	-3.64/-21.63	-38.03/-58.68	3.95/4.63	4.18/4.54	3.82/4.72
EBIT Margin	2.76/-4.71	-1.73/-30.34	-57.68/-1468.64	5.07/5.52	4.99/5.48	5.17/5.75
Stock turnover	3.16/5.11	2.75/4.31	2.57/15.37	5.61/9.34	5.75/10.49	5.86/10.74
Interest cover	1.11/7.00	-0.78/-1.32	-9.49/-26.15	6.90/31.17	6.97/30.27	7.82/39.82
Credit period	97.94/131.83	114.01/232.88	230.94/2611.99	46.86/53.32	42.93/49.07	43.03/48.82
Debt ratio	93.14/94.28	97.45/114.32	138.58/182.67	66.88/65.77	64.90/64.48	63.64/63.14

Table 7. *Correlation coefficients for the variables used in individual years (2011/2012/2013).*

	Current ratio	Cash flow liquidity	Net working capital	Return on assets	EBIT Margin	Stock turnover	Interest cover	Credit period	Debt ratio
Current ratio	1/1/1	.75/./.21/-.15	.12/.21/.12	.19/.09/.02	.18/.05/.00	.01/.01/-.02	.07/.07/.15	-.21/-.08/.00	-.36/-.26/-.26
Cash flow liquidity		1/1/1	.06/.06/.02	.48/.54/.48	.46/.20/.02	.03/.01/-.02	.02/.29/.08	-.22/-.11/-.02	-.30/-.44/-.37
Net working capital			1/1/1	.05/.05/.04	.08/.03/.01	-.02/-.02/-.04	.01/.01/.00	-.08/-.04/-.01	-.13/-.09/-.06
Return on assets				1/1/1	.20/.09/0.06	.04/.00/-.07	.35/.23/.21	-.26/-.08/-.06	-.35/-.48/-.64
EBIT Margin					1/1/1	.02/.01/.01	.27/.08/.02	-.20/-.29/-.48	-.34/-.08/-.06
Stock turnover						1/1/1	.02/.00/.00	-.10/-.04/-.01	.00/.02/.06
Interest cover							1/1/1	-.08/-.04/-.02	-.21/-.23/-.20
Credit period								1/1/1	.24/.13/.07
Debt ratio									1/1/1

SORT 47 (1) January-June 2023, 151-186 DOI: 10.57645/20.8080.02.4

Data wrangling, computational burden, automation, robustness and accuracy in ecological inference forecasting of R×C tables

Jose M. Pavía[1] and Rafael Romero[2]

Abstract

This paper assesses the two current major alternatives for ecological inference, based on a multinomial-Dirichlet Bayesian model and on mathematical programming. Their performance is evaluated in a database made up of almost 2000 real datasets for which the actual cross-distributions are known. The analysis reveals both approaches as complementarity, each one of them performing better in a different area of the simplex space, although with Bayesian solutions deteriorating when the amount of information is scarce. After offering some guidelines regarding the appropriate contexts for employing each one of the algorithms, we conclude with some ideas for exploiting their complementarities.

MSC: *90C05, 62F15, 62Q05, 05A17, 91-08, 05-04.*

Keywords: *Ecological inference; Voter transitions; US voting rights; two-way contingency tables; ei.MD.bayes; lphom; R-packages.*

1. Introduction

Ecological inference forecasting aims to estimate the inner-cells values of a set of related R×C contingency tables when only the margins are known. Ecological inference is a particular instance of cross-level inference. In ecological inference, the objective is to infer individual-level behavior from aggregate-level (i.e., ecological) data when individual-level data are not available. This outlines one of the more conspicuous and

[1] GIPEyOP, Universitat de Valencia, Valencia (Spain), email: pavia@uv.es ORCID: https://orcid.org/0000-0002-0129-726X

[2] Universidad Politécnica de Valencia, Valencia (Spain), email: rromero@eio.upv.es

Received: September 2022
Accepted: April 2023

long-standing problems of social sciences present in many disciplines, from marketing and epidemiology to sociology and political science, and encompassing geography, economics and quantitative history (King, 1997; Petropoulos et al., 2022). In ecological inference, the problem arises because information is lost when aggregating across individuals, the fundamental challenge being that many different possible relationships at the individual level can produce the same observations at the aggregate level.

Despite the dangers of cross-level inferences being widely acknowledged, arising from the so-called group or ecological fallacy (e.g., Allport, 1924; Robinson, 1950) and the Simpson paradox (e.g., Gehlke and Biehl, 1934; Simpson, 1951), the solutions promised by this approach soon attracted the interest of researchers, mainly within the discipline of political science (Ogburn and Goltra, 1919; Gosnell and Gill, 1935). A particularly relevant instance of this problem arises when the focus is on estimating/forecasting the inner-cells values of a set of related R$\times$C contingency tables when only the margins are known. For example, finding out from the data available on a set of voting units (e.g., counties or precincts) how different people (grouped, for instance, according to their religion: Catholics, Protestants, Muslims, agnostics, ...) split their votes among different candidates, or estimating the vote transfers between two elections. Focusing on the second example, the objective is to ascertain the cross-tabulated distribution of votes in each unit and in the whole electoral space by just using the sets of votes recorded in the units in the two elections (the margins of the tables).

The fundamental challenge of the ecological inference forecasting problem lies in the fact that there are a multitude of ways to determine the interior cell counts of a table with the same aggregated margins, and this indeterminacy cannot be solved collecting data from more units (Manski, 2007; Greiner and Quinn, 2009; Forcina and Pellegrino, 2019). To disentangle this, a basic assumption of similarity (and, sometimes, the use of covariates) is routinely considered. The aim of this paper is to assess in terms of accuracy, robustness and simplicity, and also considering the features of computational burden, automation and data wrangling requirements, the two main alternatives for ecological inference forecasting available in the R packages `eiPack` (Lau, Moore and Kellermann, 2020) and `lphom` (Pavía and Romero, 2021).

Klima et al. (2016) and Plescia and De Sio (2018), working independently and after analyzing the main methods developed up to that moment, conclude that the algorithm programmed in the `ei.MD.bayes` function of the `eiPack` package is the one that generates the best solutions. However, Romero et al. (2020) and Pavía and Romero (2022) have recently proposed three new algorithms (`lphom`, `tslphom` and `nslphom`), available in the `lphom` package (Pavía and Romero, 2021), whose performance seems to exceed, at least in certain circumstances, the estimates achieved with `ei.MD.bayes`. Romero et al. (2020) and Romero and Pavía (2021) report, when studying the vote transfers between the first and second rounds of the 2017 French presidential elections, that `ei.MD.bayes` produces unusable solutions when working with limited voting units.

Specifically, Romero and Pavía (2021) find when working with outcomes at the regional level (13 voting units) and with outcomes at the department level (108 units)

that `ei.MD.bayes` generates solutions without socio-political sense. They obtain this result both when using the default options of `ei.MD.bayes` and when tuning the parameters of the function. Only when working with outcomes at the district level (577 units), and after tuning the model parameters (incurring significant data analysis and computational costs), are they able to achieve satisfactory solutions. These findings contrast, on the one hand, with the excellent solutions that `ei.MD.bayes` provides, using its default options, for the dataset (212 units) included in the `eiPack` package and with the conclusions reported in Klima et al. (2016) and Plescia and De Sio (2018) and, on the other hand, with the satisfactory solutions that are achieved, in a few seconds and without specification costs, using the default options of the `lphom` functions. Therefore, a broad and systematic study of comparison is needed between the functions of both packages to determine the empirical strengths and weaknesses of each algorithm and the circumstances in which each of them will generate better estimates.

Although a significant part of the studies of this nature use simulated datasets to assess the quality of the estimates (e.g., Ferree, 2004; Greiner and Quinn, 2010; Klima et al., 2016; Klein, 2019; Klima et al., 2019; Martín, 2020; Barreto et al, 2022) since the data of interest in real situations is usually unknown (indeed, this is the purpose of the different procedures), in this research we use real data for the assessments. This is in line with Plescia and De Sio (2018) and Pavía and Romero (2022) and is the approach recommended by Collingwood et al. (2016, p. 93), who "argue that real election data should be considered in a side-by-side comparison". In particular, the performance of the different algorithms is evaluated by exploiting the data from a singular database made up of almost 500 elections for which the current cross-distribution of votes in the entire electoral space is known. This database includes all the elections analyzed in Plescia and De Sio (2018) and Pavía and Romero (2022).

The assessment of the algorithms will not only focus on evaluating their accuracy in predicting the cross distributions but also on other considerations such as their data wrangling and specification requirements. On the one hand, the procedure implemented in `ei.MD.bayes` is a complex procedure, based on Markov chain Monte Carlo (MCMC) methods, that (i) demands the specification and tuning of a large number of parameters (among them, a priori distributions with their hyperparameters, the number of initial iterations to be discarded, the length of the chains or the jump between accepted values in each chain) and (ii) requires, before using the function, an intensive data pre-processing to guarantee the congruence between the marginal distributions of rows and columns of each table. On the other hand, the procedures implemented in `lphom`, based on mathematical programming, can negotiate different scenarios in terms of (lack of) congruence between marginal distributions and only require, in the `nslphom` algorithm, specification of the number of iterations. All these issues must be weighed up when choosing an algorithm to solve a problem.

Given that in real situations the inner-cells of the contingency tables are generally unknown – at most, we can check the solutions for their plausibility but not the quality (accuracy) of the predictions – we also evaluate the robustness and sensitivity of the

different algorithms in either more stressful or simpler scenarios. Starting from the observed database composed of 493 elections, we construct new sets of electoral results by aggregating voting units and/or voting options. This will allow the scenarios under analysis to be increased and the algorithms to be evaluated in new situations, where the problem is simplified (with fewer cells in the transfer matrices) and/or with less data available (with fewer voting unit observations). In total, using real data at all times, we analyze the equivalent of 1972 elections.

The rest of the document is structured as follows. The second section details the characteristics of the methods implemented in both functions. The third section is dedicated to data. The fourth section compares and analyses the results attained after applying `ei.MD.bayes`, with different specifications, and the `lphom` algorithms to the initial datasets corresponding to the 493 elections. The fifth section explores the robustness and sensitivity of the estimates in the new scenarios, created from the base data. Section 6 reviews the analysis and, by pooling the results of all the datasets, looks at, among other issues, the features that affect the accuracy of the estimates in both approaches. Finally, Section 7 summarizes the findings, states some recommendations and suggests directions for further research. This paper is complement with two files with Supplementary Material.

2. The methods

In the ecological inference forecasting problem, the units of analysis are contingency tables with observed row and column marginals and the objective is to estimate the unobserved internal cells for each unit (and/or for the aggregation of all the tables). Mathematically, denoting by $i = 1,2,...,I$ the index for the units, $j = 1,2,...,R$ the index for the rows and $k = 1,2,...,C$ the index for the columns (where I, R and C represent, respectively, the number of units, rows and columns), the problem can be stated, as expressed in Table 1, as one of estimating the (red) values $N_{jki}\ \forall i,j,k$, given their row and column aggregations: $N_{j.i} = \sum_k N_{jki}$ and $N_{.ki} = \sum_j N_{jki}$ (where $N_{..i} = \sum_{jk} N_{jki} = \sum_j N_{j.i} = \sum_k N_{.ki}$).

Table 1. *A typical $R \times C$ unit in ecological inference. Red quantities are the unobserved counts.*

	col_1	...	col_k	...	col_C	
row_1	N_{11i}	...	N_{1ki}	...	N_{1Ci}	$N_{1.i}$
...	...	...	...	...	...	...
row_j	N_{j1i}	...	N_{jki}	...	N_{jCi}	$N_{j.i}$
...	...	...	...	...	...	...
row_R	N_{R1i}	...	N_{Rki}	...	N_{RCi}	$N_{R.i}$
	$N_{.1i}$	...	$N_{.ki}$	...	$N_{.Ci}$	$N_{..i}$

Many algorithms for solving the ecological inference forecasting problem can be found in the literature. In this research, the estimates obtained from two procedures with

different philosophical and mathematical substrates are compared: on the one hand, the three algorithms implemented in the `lphom` package (Pavía and Romero, 2021) and, on the other hand, several specifications of the procedure available in the `ei.MD.bayes` function of the `eiPack` package (Lau et al., 2020). The first algorithms are based on mathematical programming, while the second procedure has its roots in Bayesian statistics. Other methods to solve this problem include the iterative version of the 2×2 model proposed by King (see King, 1997; Imai, King and Lau, 2008; Collingwood et al., 2016; Choirat et al., 2017), the aggregated compound multinomial model proposed by Brown and Payne (1986) or the generalization of the Goodman regression method (see Goodman, 1953, 1959; Collingwood et al., 2016).

Despite the different foundations of the various procedures, they all rely on the same information sources and basic assumptions to obtain their estimates. All of them exclusively use the information contained in the margins of the tables and assume a hypothesis of similar behavior between different units to overcome the problems of identifiability and indeterminacy. In particular, `lphom` assumes small distances across units among p^i_{jk} and also with p_{jk} and `ei.MD.bayes` considers that, conditional on the row, j, all the p^i_{jk} of the different units are realizations of a common probability distribution, where $p^i_{jk} = N_{jki}/N_{j.i}$ and $p_{jk} = \sum_i N_{jki}/\sum_i N_{j.i}$ are, respectively, the (unknown) unit and global cell fractions. Both procedures also impose (explicitly `lphom` and implicitly `ei.MD.bayes`) the restrictions that are derived from the available information. The unit cell fractions, p^i_{jk}, that both approximations estimate must be compatible with the marginals of each unit and of the set of tables.

2.1. The model in ei.MD.bayes

The procedure implemented in the `ei.MD.bayes` function uses a method based on a hierarchical Multinomial-Dirichlet model initially proposed for 2×2 tables by King, Rosen and Tanner (1999) and later generalized for R×C tables by Rosen et al. (2001). Specifically, denoting the row marginal and the column marginal fractions of unit i by, respectively, $X_{ji} = N_{j.i}/N_{..i}$ and $T_{ki} = N_{.ki}/N_{..i}$, the hierarchical Multinomial-Dirichlet model, without covariates, proposed by Rosen et al. (2001) assumes, for the first level of the hierarchy, that the vector of column marginal counts in unit i follows a Multinomial distribution of the form:

$$(N_{.1i},...,N_{.ki},...,N_{.Ci}) \sim \text{Multinomial}(N_{..i}, \sum_{j=1}^{R} p^i_{j1}X_{ji}, ..., \sum_{j=1}^{R} p^i_{jk}X_{ji}, ..., \sum_{j=1}^{R} p^i_{jC}X_{ji})$$

and, for the second level of the hierarchy, that the vector of cell fractions for row j ($j = 1,\ldots,R$) in unit i ($i = 1,\ldots,I$) follows a Dirichlet distribution with C parameters, constant across units:

$$(p^i_{j1},\ldots,p^i_{jk},\ldots,p^i_{jC}) \sim \text{Dirichlet}(\alpha_{j1},\ldots,\alpha_{jk},\ldots,\alpha_{jC})$$

where the prior on each α_{jk} is assumed to be:

$$\alpha_{jk} \sim \text{Gamma}(\lambda_1, \lambda_2)$$

The first level of the hierarchy introduces the information of the margins by modelling, conditional on the observed row totals, the observed column totals as multinomial distributions independent across units. The second level of the hierarchy enables the borrowing of strength across the estimates of the (unobserved) row-cell proportions/fractions of different units by modelling them as Dirichlet distributions independent across rows and conditional independent across units. The third level of the hierarchy considers a fairly non-informative distribution for the Dirichlet parameters. The hierarchical model not only increases efficiency (decreases variation) of the estimates by borrowing statistical strength across units, but it also makes it possible to obtain estimates of the unobserved quantities p^i_{jk}.

This hierarchical Bayesian model is fit by `ei.MD.bayes` using a Metropolis-within-Gibbs algorithm (Robert and Casella, 2004). Conducting an analysis employing this model involves two steps: first, calibrating priors and tuning parameters used for Metropolis-Hastings sampling and, second, generating proper MCMC draws. This requires analysts highly trained in Bayesian statistics since, in addition to the need to tune a large number of parameters, assessing convergence of MCMC chains tends to be difficult is this setting (Rosen et al., 2001; Lau, Moore, and Kellermann, 2007): sometimes the scarce information available in the margins of the tables (i.e., regarding p^i_{jk} bounds) can lead to extremely slow mixing of MCMC chains. Furthermore, when the number of units is scarce and all the margins of the unit tables are sufficiently populated, some substantive knowledge of the phenomenon under study is also required to properly customize prior hyperparameters. As Wakefield (2004) notes, the inherent problems of identifiability and indeterminacy that characterizes ecological inference is likely to lead to solutions sensitive to the choice of prior so, as Lau et al. (2007, p. 46) recommend, "[u]sers should experiment with different assumptions about the prior distribution of the upper-level parameters in order to gauge the robustness of their inference". It is also necessary to properly set issues such as the length of the burn-in period, the thinning parameter and the total length of the chains. It is essential to generate enough iterations for the Markov Chain to converge, as only if a convergence occurs can the samples from a Markov Chain be used in a Monte Carlo integration.

2.2. The model in lphom

The methods included in `lphom`, acronym for "**L**inear **P**rogram model based on the **HOM**ogeneity hypothesis", estimate the p^i_{jk} by solving two sequential linear programs that, conforming to the observed marginal counts, minimizes the L_1 distance of the cell fractions across units. The `nslphom` algorithm (Pavía and Romero, 2022) is an iterative procedure that yields the `lphom` and the `tslphom` solutions as by-products. In its simplest specification, `nslphom` uses equations (1) to (15) to attain its solution. In its

step zero, the algorithm solves the basic `lphom` system (Romero et al., 2020) defined by equations (1) to (5).

$$p_{jk} \geq 0 \quad \text{for } j = 1,\ldots,R,\ k = 1,\ldots,C \tag{1}$$

$$\sum_{k=1}^{C} p_{jk} = 1 \quad \text{for } j = 1,\ldots,R \tag{2}$$

$$\sum_{j=1}^{R} \left(\sum_{i=1}^{I} N_{j\cdot i} \right) p_{jk} = \sum_{i=1}^{I} N_{\cdot ki} \quad \text{for } k = 1,\ldots,C \tag{3}$$

$$e_{ik} = N_{\cdot ki} - \sum_{j=1}^{R} N_{j\cdot i} p_{jk} \quad \text{for } k = 1,\ldots,C,\ i = 1,\ldots,I \tag{4}$$

$$\textit{minimize} \sum_{i,k} |e_{ik}| \tag{5}$$

This step zero produces an initial solution matrix $\hat{\mathbf{P}}_0 = \left[{}_0\hat{p}_{jk} \right]$ of the matrix, $\mathbf{P} = \left[p_{jk} \right]$, of global cell fractions that is used to start the iterative process that characterizes `nslphom`. In the next steps, for $l = 1, ..., ns$ (where ns is the number of steps), the algorithm generates estimates of the unit cell fractions, p^i_{jk}, and the global cell fractions, p_{jk}, by recursively updating the ${}_l\hat{p}_{jk}$ estimates and solving the two sequential systems defined by expressions (6) to (13).

$$p^i_{jk} \geq 0 \quad \text{for } j = 1,\ldots,R,\ k = 1,\ldots,C,\ i = 1,\ldots,I \tag{6}$$

$$\sum_{k=1}^{C} p^i_{jk} = 1 \quad \text{for } j = 1,\ldots,R,\ i = 1,\ldots,I \tag{7}$$

$$\sum_{j=1}^{R} N_{j\cdot i} p^i_{jk} = N_{\cdot ki} \quad \text{for } k = 1,\ldots,C,\ i = 1,\ldots,I \tag{8}$$

$$\varepsilon^i_{jk} = ({}_{l-1}\hat{p}_{jk} - p^i_{jk}) N_{j\cdot i} \quad \text{for } j = 1,\ldots,R,\ k = 1,\ldots,C,\ i = 1,\ldots,I \tag{9}$$

$$\textit{minimize } Z = \sum_{j,k} |\varepsilon^i_{jk}| \quad \text{for } i = 1,\ldots,I \tag{10}$$

$$Z = \sum_{j,k} |\varepsilon^i_{jk}| \quad \text{for } i = 1,\ldots,I \tag{11}$$

$$p^i_{jk} = ({}_{l-1}\hat{p}_{jk} + \delta^i_{jk}) \quad \text{for } j = 1,\ldots,R,\ k = 1,\ldots,C,\ i = 1,\ldots,I \tag{12}$$

$$\textit{minimize} \sum_{j,k} |\delta^i_{jk}| \quad \text{for } i = 1,\ldots,I \tag{13}$$

where ${}_l\hat{p}_{jk}$ is computed by equation (14) using the l-step solutions $\hat{p}^i_{jk}(l)$ attained after solving equations (6)-(13).

$${}_l\hat{p}_{jk} = \sum_{i=1}^{I} \hat{p}^i_{jk}(l) N_{j\cdot i} / \sum_{i=1}^{I} N_{j\cdot i} \quad \text{for } j = 1,\ldots,R,\ k = 1,\ldots,C \tag{14}$$

During the iterative process, the statistic defined by equation (15), which measures the aggregate distance to homogeneity of the recursive solutions, is also computed. This statistic is utilized to determine the `nslphom` solution, which corresponds to the iteration l^* minimizing (15) .

$$HET_l = 100 \cdot \frac{0.5 \sum_{ijk} \hat{p}^i_{jk}(l) N_{j\cdot i} - {}_l\hat{p}_{jk} N_{j\cdot i}}{\sum_{ij} N_{j\cdot i}} \tag{15}$$

Once the iterative process has finished, we have three solutions: the `lphom` solution, which corresponds to the step zero solution, the `tslphom` solution, which corresponds to the solution attained in step one and, finally, the solution corresponding to step l^*, which is the `nslphom` solution. Note that the `lphom` solution only provides estimates for the inner-cells of the global table. The above algorithm is quite automatic with only one parameter to tune: the number of steps, *ns*. According to Pavía and Romero (2022), the minimum of equation (15) is usually reached after very few steps. Indeed, the default option of the `nslphom` function considers only ten steps.

3. The data

Given the secret nature of voting, internal cell counts of global and unit tables are mostly unobserved. Sometimes, however, they are available, as when voters cast ballots with several votes in the same ballot and they are counted and published jointly. This is (partially) the case of the New Zealand general elections since 2002 and of the 2007 Scottish Parliamentary election, where a mixed-member election system is employed. In these elections, voters cast two independent votes – one for a list (usually a party list) and another for a local candidate – and the electoral authorities publish/published party-candidate cross-tabulations at district level and marginal results at polling station level. This provides a unique opportunity to assess algorithms by comparing actual observed global cross-tables with forecasted ecological tables. In each district, the `ei.MD.bayes` and `nslphom` functions can be run to forecast the internal cell counts (or fractions) of the district table using as inputs the marginal results at polling station level, to afterwards compare forecasts and actual observed values.

Specifically, we collected 493 datasets composed of marginal polling stations' results and party-candidate cross tables corresponding to the same number of elections (districts): 420 datasets came from the 2002, 2005, 2008, 2011, 2014 and 2017 New Zealand general elections and 73 datasets from the 2007 Scottish Parliament election. In the case of New Zealand, the raw files of the cross-distributions of votes at district level (with parties by rows and candidates by columns) and of the marginal distributions of votes at polling station level were downloaded from the official web page of the electoral commission of New Zealand (www.electionresults.org.nz). In the case of Scotland, the authors gained access to the data via personal communication with Carolina Plescia,

who had downloaded the raw files from the Scotland Electoral Office website in 2011. The Scottish data are no longer available on that site.

Before using the data, every election-district dataset is checked for internal consistency and pre-processed in order to guarantee that the accounting equalities $\sum_j N_{j\cdot i} = \sum_k N_{\cdot ki}$ (for $i = 1,...,I$) and $\sum_i N_{j\cdot i} = \sum_k N_{jk\cdot}$ and $\sum_i N_{\cdot ki} = \sum_j N_{jk\cdot}$ (for $j = 1,...,R$ and $k = 1,...,C$) hold in each dataset for, respectively, each polling station (voting unit) and the whole district, where $N_{jk\cdot}$ $(= \sum_i N_{jki})$ are the internal cell counts (observed in these datasets) of the district tables.

In the case of the New Zealand datasets, we have removed: (i) the rows with all their values being zero or non-available in the parties' and candidates' files; and (ii) the row corresponding to the polling unit identified as "Votes allowed for party only" in the parties' files and, equally, the corresponding column ("Party vote only") in the cross-distribution files. The second group of deletions was performed because the voting unit "Votes allowed for party only" has no equivalent in the candidates' files. In addition to these general pre-processing tasks, we merged the voting units identified as "Voting places where less than 6 votes were taken" (row 100) and "Ordinary votes before polling day" (row 101) in the party and candidate files of the 43rd district (Rangitikei) of the 2014 election. We did this to solve a mismatch between both files as the values in their 100th and 101st rows were, respectively, 3 and 2 and 8465 and 8466.

Finally, before starting any analysis and as is common practice when forecasting real tables (e.g., Klima et al., 2016; Plescia and De Sio, 2018; Klein, 2019; Pavía and Aybar, 2020; Pavía and Romero, 2022), we merged very small electoral options. In each dataset, those parties or candidates which individually did not reach at least 3% of the election-district vote were grouped in the option 'Others'. Hereinafter, we call this set of datasets the reference database. Table 2 offers some summary statistics of this database, with more details available in Pavía (2022).

As can be seen in Table 2, we have some variety in terms of the features in the datasets collected. In particular, looking at the last two columns of Table 2, we see that our database also presents an interesting diversity in terms of voters' distribution among cells within rows. And this despite our cross-tables coming from ticket-splitting in concurrent elections, where more cell fractions close to one (zero) are routinely recorded than in other contexts, such as in demographic voting. This, undoubtedly, enriches the analyses by allowing the algorithms to be evaluated in different contexts. Indeed, according to Park, Hanmer and Biggers (2014), gauging the accuracy of ecological inference procedures across different contexts adds robustness to the conclusions, particularly for studying what happens when the across-unit variance varies and/or when the number of units is small.

According to Wakefield (2004), smaller areas are preferable (i.e., voting units with a small number of voters) because it reduces the possibility of ecological bias and, likewise, it is also better to have very little within-area variability among row proportions because this leads to accurate estimates of fractions. Nevertheless, Romero and Pavía (2021) advocate studying the behavior of both algorithms when the number of units ob-

Table 2. *Summary of some features of the datasets used to evaluate the algorithms.*

Country	Year	Elections (datasets)	Average number of voting units (min-max)	voters by units[1] (min-max)	parties (min-max)	candidates (min-max)	large p_{jk} fractions[2]	% voters in large p_{jk}[3]
NZ	2002	69	$83.2_{(30-651)}$	$554.6_{(24.5-1075.5)}$	$7.0_{(5-8)}$	$5.7_{(5-8)}$	$1.2_{(0-2)}$	$36.0_{(0.0-65.3)}$
NZ	2005	69	$81.8_{(29-698)}$	$634.5_{(28.3-1194.0)}$	$5.2_{(4-7)}$	$4.5_{(3-6)}$	$1.4_{(0-2)}$	$50.5_{(0.0-77.0)}$
SCO	2007	73	$70.2_{(22-103)}$	$411.6_{(346.3-547.1)}$	$6.0_{(5-8)}$	$5.9_{(5-8)}$	$2.6_{(0-4)}$	$59.1_{(0.0-80.5)}$
NZ	2008	70	$84.1_{(32-686)}$	$614.6_{(28.7-1094.8)}$	$5.4_{(4-6)}$	$4.4_{(3-6)}$	$1.7_{(0-3)}$	$52.5_{(0.0-80.7)}$
NZ	2011	70	$85.7_{(32-644)}$	$555.0_{(27.2-1068.0)}$	$5.6_{(4-7)}$	$4.7_{(4-6)}$	$1.4_{(0-2)}$	$49.7_{(0.0-73.5)}$
NZ	2014	71	$81.2_{(31-620)}$	$617.0_{(32.6-1124.2)}$	$5.9_{(5-7)}$	$4.7_{(3-6)}$	$1.5_{(0-3)}$	$49.9_{(0.0-73.9)}$
NZ	2017	71	$101.9_{(41-705)}$	$487.7_{(33.2-1012.7)}$	$5.2_{(4-7)}$	$4.8_{(3-6)}$	$1.3_{(0-2)}$	$47.3_{(0.0-77.9)}$
Total	–	493	$84.0_{(22-705)}$	$552.2_{(24.5-1194.0)}$	$5.8_{(4-8)}$	$4.9_{(3-8)}$	$1.6_{(0-4)}$	$49.4_{(0.0-80.7)}$

Source: compiled by the authors using data from the New Zealand (NZ) electoral commission and the Scotland (SCO) Electoral Office.

[1] These averages correspond to averages of averages. First, the average number of voters per voting unit $\sum_i N_{.i}/I$ is computed for each dataset and then the average of these averages is calculated for each year.

[2] A p_{jk} is considered a large fraction when it is higher than 0.80.

[3] The percentage of voters located in cells with large fractions, $p_{jk}(> 0.80)$, in each election/dataset is computed as $100\sum_{(j,k)\in L} N_{jk.}/\sum_i N_{..i}$, where $L = \{(j,k) : p_{jk} > 0.80\}$.

served is small, since this reduces the costs of data wrangling and would help to answer the question of whether they could be used immediately after an election, a time when the results are usually available at a high level of aggregation, for a small number of units. Thus, in order to increase the number of analysis scenarios, we build new datasets by merging voting units and/or voting options (parties and candidates). On the one hand, reducing the number of units adds difficulty to the problem, by reducing the amount of information available. On the other hand, reducing the number of voting options simplifies the problem, by decreasing the number of unknowns. In general, both operations contract the across-unit variance.

We derive three new datasets from each baseline dataset by (i) reducing the number of voting units, (ii) reducing the number of cells in the tables (the number of parties and candidates), and (iii) reducing both the number of units and the number of cells. More specifically, the initial number of units of each dataset is reduced by randomly grouping the units into a random number of groups, uniformly selected between 10 and 20, and merging them. The number of parties and candidates is reduced by adding to either the row or column voting option Others, respectively, the votes of those parties or candidates that did not reach a minimum of 20% of the votes. The random merging of units in scenarios (i) and (iii) have been performed in an independent fashion in order to induce more variability in the constructed database. After all these operations, we are ready to analyze real data equivalent to the 1972 elections.

4. An initial comparison of procedures

This section, focused on accuracy, assesses the solutions achieved after applying `ei.MD.bayes` (with different specifications), `lphom`, `tslphom` and `nslphom` to the

reference database of the 493 datasets that made up our collected data before performing the processes of merging of units and/or cells described in the last paragraph of the previous section. As a rule, and starting point, we have considered the default options of the functions, given that these are usually the specifications most utilized by users. These simplify their decision-making processes, reduce their operational costs and favor automation. In the case of `ei.MD.bayes`, we consider three different specifications, which we label as `ei_default`, `ei_auto`, and `ei_manual`.

The `ei_default` solutions correspond to the use of `ei.MD.bayes` with default options. A solution based on MCMC, however, requires convergence to the equilibrium distribution of the Markov chains to be reliable. Unfortunately, this is not attained as a rule with our data when `ei.MD.bayes` is employed with default options. The arguments of the function, therefore, have to be tuned to generate convergent chains. The `eiPack` package also includes a function, `tuneMD`, with "an adaptive algorithm to generate tuning parameters for the MCMC algorithm implemented in `ei.MD.bayes`" (Lau et al., 2020). So, as a second specification for `ei.MD.bayes`, we implement a two-step strategy in which firstly `tuneMD` is employed with default options to afterwards apply `ei.MD.bayes` with its `tune.list` argument equal to the output generated by `tuneMD`. This does not solve the lack of convergence, however. This should not come as a surprise since, as the `ei.MD.bayes` help file advises: most problems will require a much larger thinning interval and a longer burn-in period than default.

At this point, it is clear that what is necessary is to manually customize the arguments of the `ei.MD.bayes` function. Unfortunately, trying to customize 493 scenarios is impractical and extremely time-consuming. An analyst needs to test several specifications for each election, plotting each of their outputs and diagnosing their convergence (Roy, 2020). Hence, as an intermediate, operative approach, we look for a specification that could work well in general. After picking at random three datasets of each block (year) of elections, we find that a two-step strategy in which firstly `tuneMD` is employed with arguments `ntunes = 10` and `totaldraws = 100000` and secondly `ei.MD.bayes` is employed with arguments `sample = 1000`, `thin = 100`, `burnin = 100000` and `tune.list` equal to the output generated by `tuneMD` reaches the convergence in all twenty-seven elections selected. We use this specification to model all the datasets. This guarantees that convergence is reached in the twenty-seven datasets checked and also, with really high probability, in the rest of datasets. We label the solutions attained by `ei.MD.bayes` using this specification `ei_manual`.

Once each algorithm is run, we gauge accuracies of solutions using as discrepancy measures the statistics EI and WPE, given by equations (16) and (17). These statistics account for the distances between the forecasted and observed matrices at the district level of, respectively, counts and proportions. The assessments of errors at the local level are unfeasible in this case as internal cell values of local units are not available in the collected datasets.

$$EI = 100 \cdot \frac{0.5 \sum_{jk} |N_{jk\cdot} - \hat{N}_{jk\cdot}|}{\sum_{jk} N_{jk\cdot}} \tag{16}$$

$$WPE = 100 \cdot \frac{\sum_{jk} N_{jk\cdot} \, |p_{jk} - \hat{p}_{jk}|}{\sum_{jk} N_{jk\cdot}} \tag{17}$$

The EI (error index) statistic is a classical measure of discrepancy (e.g., Thomsen, 1987; Klima et al., 2016; Romero et al., 2020) that quantifies the distance between matrices of counts. In our case, it accounts for the percentage of votes wrongly assigned, i.e., the minimum number of votes that should be moved among cells of the forecasted matrix to accomplish a perfect fit. Multiplication by 0.5 is done to prevent counting every incorrectly allocated vote twice. This coefficient varies between 0, when the actual and the forecasted matrices coincide, and 100, when no single vote is correctly assigned. The WPE statistic (proposed in Pavía and Romero, 2022) measures the weighted average distance between the actual p_{jk} and the estimated $\hat{p}_{jk}$ proportions, using as weights the corresponding actual counts. This statistic ponders more the discrepancies associated with the most relevant proportions and also ranges between 0, when $\mathbf{P}$ and $\hat{\mathbf{P}}$ match, and 100, when no vote is correctly assigned. EI and WPE are closely correlated.

Table 3 synthesizes the discrepancies, measured using the EI and WPE statistics, between the actual matrices and the solutions attained after applying `lphom`, `tslphom`, `nslphom` and `ei.MD.bayes` (with the three specifications detailed above) to the datasets of the reference database. The table presents, by group of elections, average figures of EI and WPE values as well as average computation times (lower panel). The upper panel of the table also offers some summary statistics of the corresponding group of elections. The elections are naturally grouped by country and year. Ultimately, all the elections of each group are related since they were held simultaneously, sharing the same general political context. The last column summarizes the results corresponding to the whole database.

Figures 1 and 2 display the same information shown in the EI and WPE panels of Table 3, but graphically. Interested readers can also consult Figure S1 of the supplementary material which displays graphically the averages times of computation (in seconds) required to reach the solutions. Several initial findings emerge analyzing Figures 1 and 2 and the numbers in Table 3. First, all the methods yield solutions superior to a random assignment. Second, as expected, the `ei_default` and `ei_auto` solutions are by far the least accurate, given their lack of convergence. They are, nevertheless, superior to a random assignment. This may seem surprising at first glance, however, despite their failure to converge, they already include the information available in the margins of the tables; an issue that limits the set of possible solutions. Third, within the `lphom` family, `nslphom` is the one that is most accurate. This confirms the conclusions reached in Pavía and Romero (2022). Fourth, both `ei_manual` and `nslphom` solutions stand out for being the most accurate, being indeed fairly good considering the magnitude of error that, according to Klima et al. (2016), is usual in these kind of problems. Fifth, the `ei_auto` and `ei_manual` specifications require much more time than the rest of the procedures to reach their solutions.

Table 3. *Summary of the performance of the algorithms in the original datasets.*

Country Year	NZ 2002	NZ 2005	SCO 2007	NZ 2008	NZ 2011	NZ 2014	NZ 2017	NZ + SCO
# of Elections	N = 69	N = 69	N = 73	N = 70	N = 70	N = 71	N = 71	N = 493
Avg. # of units	$\overline{I}$ = 83.2	$\overline{I}$= 81.8	$\overline{I}$= 70.2	$\overline{I}$= 84.1	$\overline{I}$= 85.7	$\overline{I}$= 81.2	$\overline{I}$= 101.9	$\overline{I}$= 84.0
Avg. # of cells	$\overline{RC}$= 39.5	$\overline{RC}$= 23.8	$\overline{RC}$= 35.2	$\overline{RC}$ = 23.4	$\overline{RC}$= 26.2	$\overline{RC}$= 27.9	$\overline{RC}$= 24.8	$\overline{RC}$= 28.7
Average of *EI* mesasures								
ei_default	22.75	27.69	48.33	31.19	29.26	32.40	34.38	32.42
ei_auto	25.20	28.96	46.85	30.89	30.17	33.18	33.93	32.85
ei_manual	10.75	8.53	23.09	8.34	7.68	7.88	6.93	10.52
nslphom	12.79	9.68	8.86	9.11	9.46	9.69	8.91	9.77
tslphom	14.80	11.09	11.00	10.88	11.50	11.66	10.91	11.68
lphom	16.88	12.29	12.92	12.22	12.99	12.95	12.20	13.20
Average of *WPE* mesasures								
ei_default	16.29	21.70	41.55	25.26	23.30	26.32	28.11	26.20
ei_auto	18.44	22.70	40.46	25.04	23.94	26.78	27.67	26.54
ei_manual	6.30	5.61	18.47	5.86	4.88	4.86	4.54	7.28
nslphom	7.90	6.09	4.80	6.09	6.26	6.55	5.67	6.18
tslphom	9.42	7.52	6.72	7.90	8.05	8.15	7.46	7.89
lphom	10.82	8.46	8.07	8.89	9.13	9.04	8.39	8.96
Average of computational burden (in secs)								
ei_default	2.08	1.23	1.33	1.14	1.55	1.48	1.52	1.48
ei_auto	958.57	573.53	603.36	531.03	724.65	692.13	722.93	690.06
ei_manual	1150.58	687.37	765.20	636.40	864.02	827.75	853.43	825.70
nslphom	5.41	5.32	5.88	5.85	5.61	5.28	6.80	5.74
tslphom	0.92	0.85	0.81	0.87	0.87	0.81	0.97	0.88
lphom	0.56	0.64	0.25	0.52	0.64	0.60	0.64	0.55

Source: compiled by the authors after applying the function `ei.MD.bayes` of the R package `eiPack` (Lau et al., 2020) with different specifications and the functions `lphom`, `tslphom` and `nslphom` of the R package `lphom` (Pavía and Romero, 2021) to the official data from the New Zealand electoral commission and the Scotland Electoral Office described in Section 3. The outcomes labelled as `ei_default`, `ei_auto` and `ei_manual` have been attained after using `ei.MD.bayes` with, respectively, (i) default options, (ii) the output of the function `tuneMD` (with default options) as `tune.list` argument and default options for the rest of its arguments and (iii) `sample = 1000`, `thin = 100`, `burnin = 100000` and the output of function `tuneMD` with `ntunes = 10` and `totaldraws = 100000` as `tune.list` argument. The computations have been performed on a desktop computer with a CPU processor Intel® Core ™ i7-4930K (6 cores) 3.40GHz and 32GB of RAM.

Looking at the outcomes of Table 3 in more detail reveals further findings. Sixth, as a rule, the performance of all methods worsen when either the number of cells grows or when the number of units decreases, but it seems that the accuracy of the `ei.MD.bayes`-based solutions suffer significantly more than the `lphom`-based solutions when the number of units decreases. Seventh, it seems that most of the time `ei_manual` produces slightly better solutions than `nslphom`, but with `nslphom` being more robust. Indeed, `nslphom` beats `ei_manual` after pooling all the elections. This, however, is a consequence of the poor performance of `ei_manual` in some Scottish datasets (see Figure S8 in the Supplementary Material) due to a lack of convergence, which in this case can be solved working with larger chains. We investigate these results further in the sections that follow.

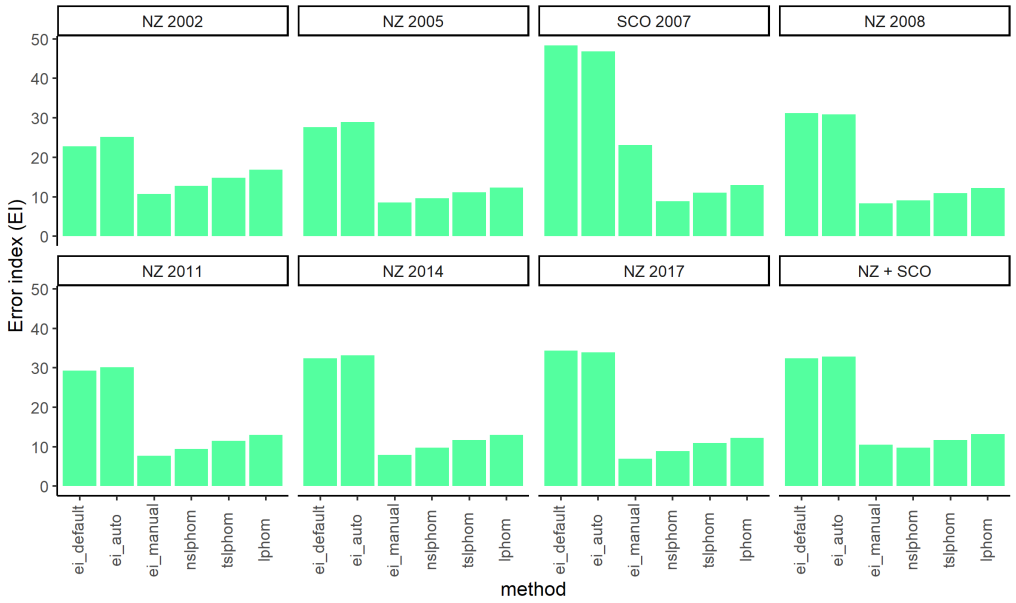

Figure 1. *Graphical representation of average values of EI error measures grouped by election and algorithm in the reference database. Individual solutions have been attained with the function* `ei.MD.bayes` *of the R package* `eiPack` *(Lau et al., 2020) using three different specifications and the functions* `lphom`*,* `tslphom` *and* `nslphom` *of the R package* `lphom` *(Pavía and Romero, 2021) with default options. Details of the specifications used when applying* `ei.MD.bayes` *can be consulted at the bottom of Table 3.*

From the above list of findings, we can gain some interesting insights. Firstly, the solutions reached using the default options of `ei.MD.bayes` are, as a rule, scarcely accurate. Despite the advantages users may find in employing functions with default options without more inquiries, this should be avoided in the case of `ei.MD.bayes`. Secondly, the default solutions of `ei.MD.bayes` can be significantly improved with some extra work by tuning all its parameters. Thirdly, the functions of the `lphom` package produce highly competitive solutions in an automatic way. Finally, the `lphom`-based solutions are, at least in these examples, reached in very few seconds.

5. Sensitivity and robustness. The effects of reducing the number of units and/or cells

The previous section evaluates `ei.MD.bayes` and `nslphom` in a set of scenarios where the relationship between the amount of information available (number of units) and the complexity of the problem (number of cells in the matrix) is considered adequate. On average, there are 2.95 voting units for each parameter to estimate when, according to Plescia and De Sio (2018, p. 673), "the literature specifies a criterion of at least two [sub]units per coefficient" for a proper forecasting of district level fractions. Although the average number of cells that we have had to estimate per election is high

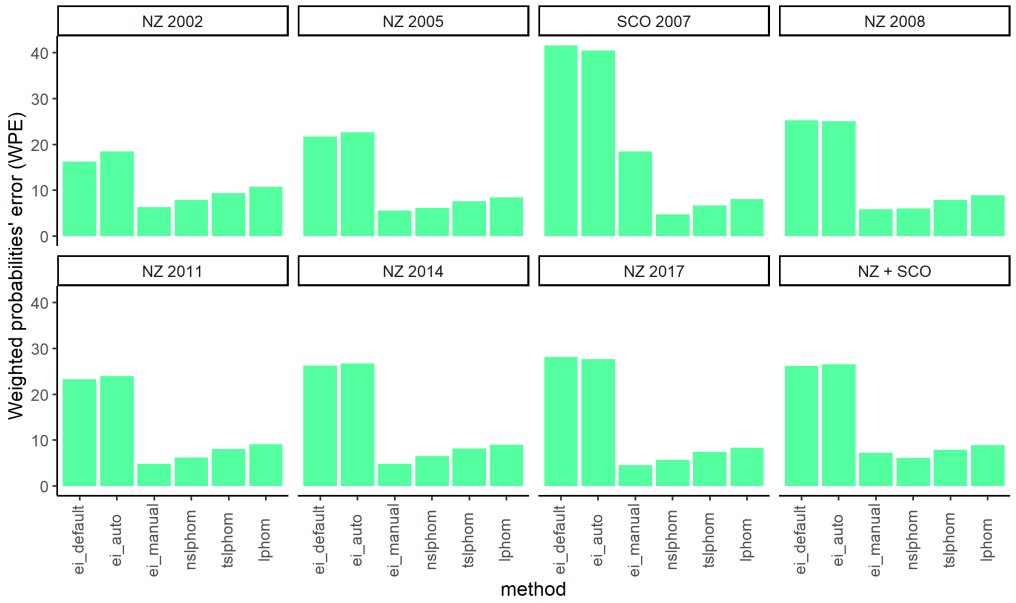

Figure 2. *Graphical representation of average values of WPE error measures grouped by election and algorithm in the reference database. Individual solutions have been attained with the function* `ei.MD.bayes` *of the R package* `eiPack` *(Lau et al., 2020) using three different specifications and the functions* `lphom`*,* `tslphom` *and* `nslphom` *of the R package* `lphom` *(Pavía and Romero, 2021) with default options. Details of the specifications used when applying* `ei.MD.bayes` *can be consulted at the bottom of Table 3.*

(28.4), so is the average number of voting units available (84), with a range that varies between a minimum of 22 and a maximum of 705, although with only 6 and 36 elections above 600 units and 200 units, respectively. Under these conditions, we get, on average, predictions of a high and similar quality, both using the `ei_manual` specification of `ei.MD.bayes` and the default options of `nslphom`. In this section, we study how the different algorithms respond when adding to the problem, by reducing the number of units, and/or through its simplification, by reducing the number of unknowns.

It is important to understand the sensitivity and robustness of the estimates when using a decreased number of units because, firstly, there are situations where obtaining more disaggregated data may be limited or even impossible (for example, in historical elections) and, secondly, because, depending on its costs in terms of accuracy, it is an option worth considering as decreasing the number of units can lead to a drastic reduction in the expenses of obtaining and handling data. It is also relevant to study how the methods behave when the number of unknowns is reduced, focusing on just the main cells. After all, the analyst, on occasions, is not interested in an overall vision of the matrix but rather in certain relevant fractions/transfers.

To answer the previous research questions, we use the three new databases derived, as stated in Section 3, from the reference database. Note that we have created three additional databases, each one also composed of 493 datasets, by just (i) grouping units

in each dataset, (ii) reducing (by aggregation) the number of cells to estimate in each dataset, and (iii) merging both, units and cells, in each dataset. In this section, we first analyze the impact of reducing the number of units, then we study the effect of reducing the number of cells and, finally, we examine the joint effect of both operations.

5.1. Effects of reducing the number of units

As in Table 3, Table S1 in the supplementary material summarizes the discrepancies measured using the *EI* and *WPE* statistics between the real matrices and the solutions attained after applying `ei.MD.bayes` (with the three specifications considered), `lphom`, `tslphom` and `nslphom` to the datasets obtained by randomly merging the observed units. Figure 3 and Figures S2 and S3 in the supplementary material present graphically the information of the different panels of Table S1. Given that the general picture drawn by *EI* and *WPE* measures is quite similar, the graphical representations corresponding to the *WPE* measures from Table S1, and the equivalent analysis in next two subsections are presented only in the supplementary material in order not to overburden this presentation.

Comparing the results of Tables 3 and S1 (Figures 1 and 3) it can be seen that, as expected, the accuracy of the solutions deteriorates as a consequence of the drastic reduction in the number of units. The impact, however, is not homogeneous in all methods. Reducing the number of units changes the order of preference between the algorithms. The solution associated with the `ei_manual` of `ei.MD.bayes` is the one that suffers the most. The mean error of this approximation is multiplied by more than two: `ei_manual` goes from having the lowest mean values for *EI* and *WPE* in almost all the election blocks to registering, in all cases, values clearly higher than those of all the solutions of the `lphom` family. Within this subset of solutions, however, the order is maintained, with the `nslphom` solutions clearly dominating those of `tslphom` and `lphom`, and this despite the fact that their relative deterioration within the subgroup is higher, with a mean increase in the error of 36%.

These findings are in line with Romero et al. (2020) and Romero and Pavía (2021) who, based on the study of the French presidential elections of 2017, noted that `ei.MD.bayes` suffers significantly when the number of units is reduced. Along the same lines, despite our best efforts, we have not found any general tuning of the parameters for `ei.MD.bayes` that works well with so few units. For example, the accuracy of the estimates does not improve even after multiplying the length of the MCMC chains by ten (with the configuration `sample = 10000`, `thin = 100` and `burnin = 1000000`). This is in contrast to the results of `nlsphom` which, with its default options, continues to generate fairly accurate solutions even in these scenarios. In light of these results, we can say that the `ei.MD.bayes`-based solutions are quite sensitive to the number of available units, quickly reducing their performance as soon as the number of units decreases and that, on the contrary, the `lphom`-based solutions are more robust. In terms of computing time, all solutions are achieved in fewer seconds.

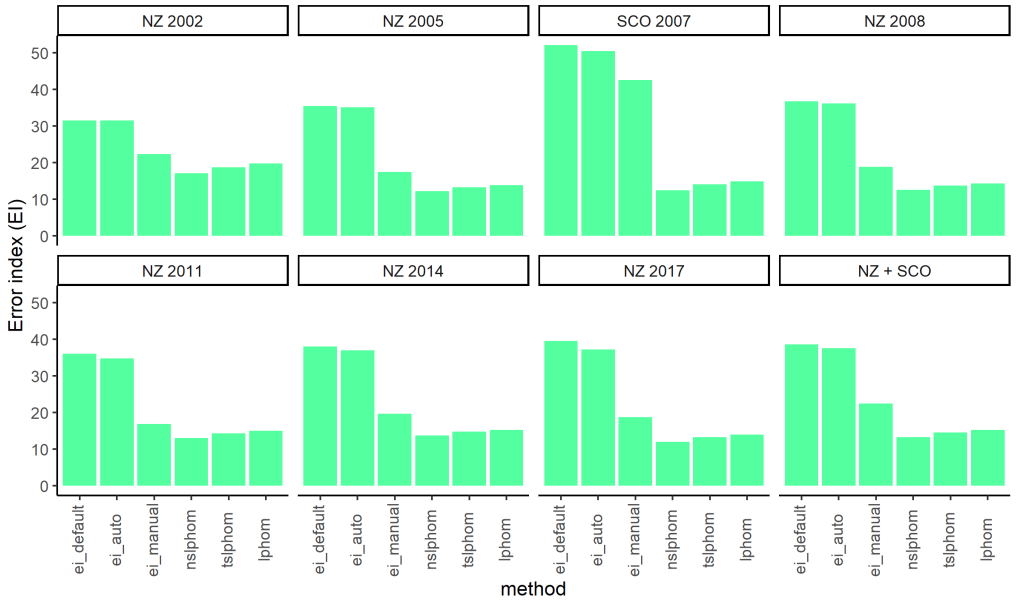

Figure 3. *Graphical representation of average values of EI error measures grouped by election and algorithm in the scenarios attained after randomly merging polling units as described in Section 3. Individual solutions have been attained with the function* `ei.MD.bayes` *of the R package* `eiPack` *(Lau et al., 2020) using three different specifications and the functions* `lphom`*,* `tslphom` *and* `nslphom` *of the R package* `lphom` *(Pavía and Romero, 2021) with default options. Details of the specifications used when applying* `ei.MD.bayes` *can be consulted at the bottom of Table 3.*

A possible explanation for the relatively worse performance of `ei.MD.bayes` in these split-ticket scenarios comes from the difficulties that its underlying (two-step) algorithm would find to move sufficiently, with so few units, the a priori row-cell fractions implied by the default values for the hyperparameters. With default options, the expected values for α_{jk} are constant by row and the expected row-cell fractions constant at $1/C$; when vote transfer matrices are characterized by having a relative large number of internal cell probabilities close to zero or one, larger than in other settings such as in racial voting applications. According to this explanation, `ei.MD.bayes` should suffer less in situations with fewer extreme fractions and/or with a lesser proportion of voters in cells with high p_{jk}. The likelihood of this explanation grows when (i) one relates the average accuracies attained in Scottish and NZ elections and their relative numbers of rows with a p_{jk} close to one (higher than 0.80) – 44.1% of rows in Scotland tables and 24.3% of rows in NZ tables have a proportion close to one – or after (ii) observing no impact in the accuracy of `ei.MD.bayes` solutions when the number of units in the `senc` dataset available in the `eiPack` package is reduced. In the `senc` dataset only 26% of voters are located in cells where $p_{jk} > 0.80$. It should be noted that with this dataset of racial voting `nslphom` neither suffers a decrease of accuracy after a reduction in the number of units.

5.2. Effects of reducing the number of cells

Table S2 in the supplementary material measures, using *EI* and *WPE*, the accuracy of the solutions achieved after running `ei.MD.bayes` (with the three specifications considered), `lphom`, `tslphom` and `nslphom` in the datasets obtained by aggregating in Others the election options not surpassing 20% of the vote. Figure 4 and Figures S3 and S4 in the supplementary material depict graphically the information of the different panels of the table.

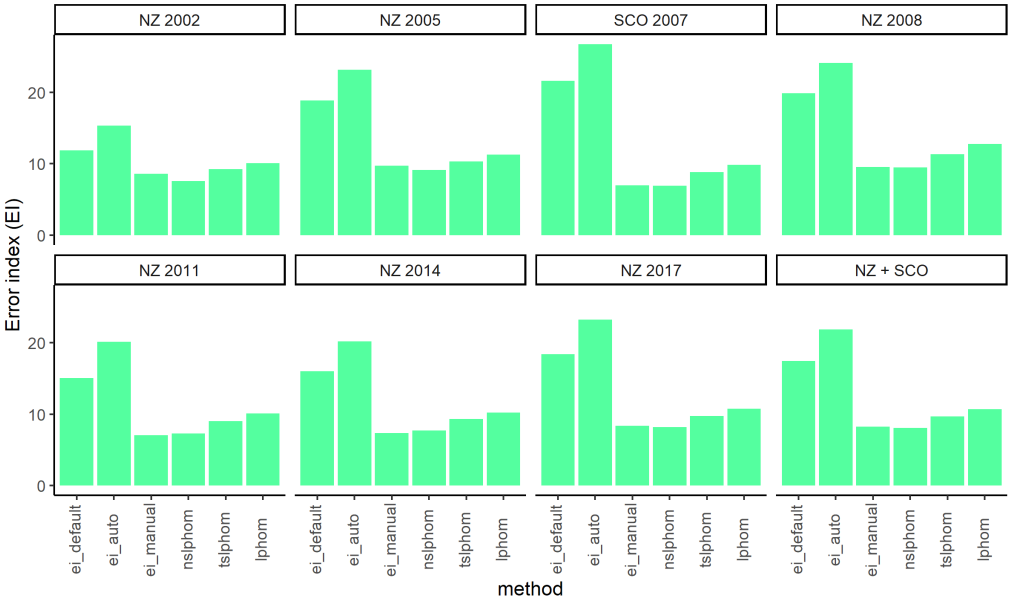

Figure 4. *Graphical representation of average values of EI error measures grouped by election and algorithm in the scenarios attained after merging in Others the election options not surpassing 20% of the vote. Individual solutions have been attained with the function* `ei.MD.bayes` *of the R package* `eiPack` *(Lau et al., 2020) using three different specifications and the functions* `lphom`, `tslphom` *and* `nslphom` *of the R package* `lphom` *(Pavía and Romero, 2021) with default options. Details of the specifications used when applying* `ei.MD.bayes` *can be consulted at the bottom of Table 3.*

Comparing the results of Tables 3 and S2 (Figures 1 and 4), it can be seen that, as expected, the accuracy of the solutions improves as a consequence of the reduction in the number of unknowns (number of cells in the transfer matrices). The general situation with respect to the baseline scenario does not change substantially. The `ei_default` and `ei_auto` specifications still do not converge, despite the reduction of unknowns, and continue to be the ones with the worst performance, while `ei_manual` and `nslphom` are the ones with the best figures, with `lphom` and `tslphom` generating highly competitive solutions. Particularly noteworthy is the fact that now the solutions for the Scottish elections with the specification `ei_manual` from `ei.MD.bayes` are significantly improved, as now all of them reach convergence. This fact means that in aggregate terms `ei_manual` is the one that most reduces its joint mean error in these

scenarios (the mean of EI goes from 10.52 to 8.22, a reduction of almost 22%). However, taking the Scottish results out of the equation, among the two main algorithms (`ei_manual` and `nslphom`), `nslphom` is revealed as the one that benefits most from the simplification of the problem. On average, it happens to be the most accurate in five of the seven election groups, when in the reference database it was only the most accurate in one of the election groups. The relative increase of rows in the target tables with a cell where $p_{jk} > 0.80$ plays, as previously discussed, against `ei.MD.bayes` as a consequence of the a priori row-cell fractions implied by the default priors for the hyperparameters. In terms of computing times, logically, costs are reduced.

5.3. Interaction effects. Effects of reducing both the number of units and cells

In subsection 5.1, we studied the effect of having fewer units and we found that solutions based on `ei.MD.bayes` suffer markedly when the number of units decreases. In subsection 5.2, we analyzed the impact of working with problems with fewer unknowns and we found that all algorithms improved their performance. In this subsection, we study what happens when both situations occur simultaneously. Table S3 in the supplementary material presents, using EI and WPE, the accuracy of the solutions reached with `ei.MD.bayes` (with the three specifications considered), `lphom`, `tslphom` and `nslphom` in the datasets obtained after reducing the number of cells and units, as stated in Section 3. Figure 5 and Figures S5 and S6 in the supplementary material show graphically the information of the different panels of Table S3.

Comparing the results of Tables 3 and S1 to S3, and the corresponding graphical representations (Figures 1 to 5), it can be seen that in this scenario the accuracies of the solutions generated by the different algorithms are at some intermediate point between the accuracies of the solutions obtained in the analyzed scenarios in subsections 5.1 and 5.2. The relative impact of both types of reductions (of data and of unknowns), however, is not homogeneous in all algorithms, at least for these datasets and with the reductions in the number of units and cells implemented. In the case of solutions based on `ei.MD.bayes`, we see that reducing the number of units has much more impact than reducing the number of cells, while in the case of solutions based on `lphom` the opposite relationship is observed, with the decrease in the number of unknowns having more relative importance. These results confirm and reinforce the conclusions reached in the previous subsections: `ei.MD.bayes` inferences are very sensitive to the data-unknowns relationship, deteriorating notably when the level of detail of the information is reduced, while `nslphom` is very robust, being more insensitive to a decrease in the amount of available data. In all cases, computing times very clearly drop.

6. A comparison of ei_manual and nslphom solutions

From the analyses carried out in sections 4 and 5, we can affirm that the `ei_manual` and `nslphom` algorithms are clearly the ones that provide, within each methodology

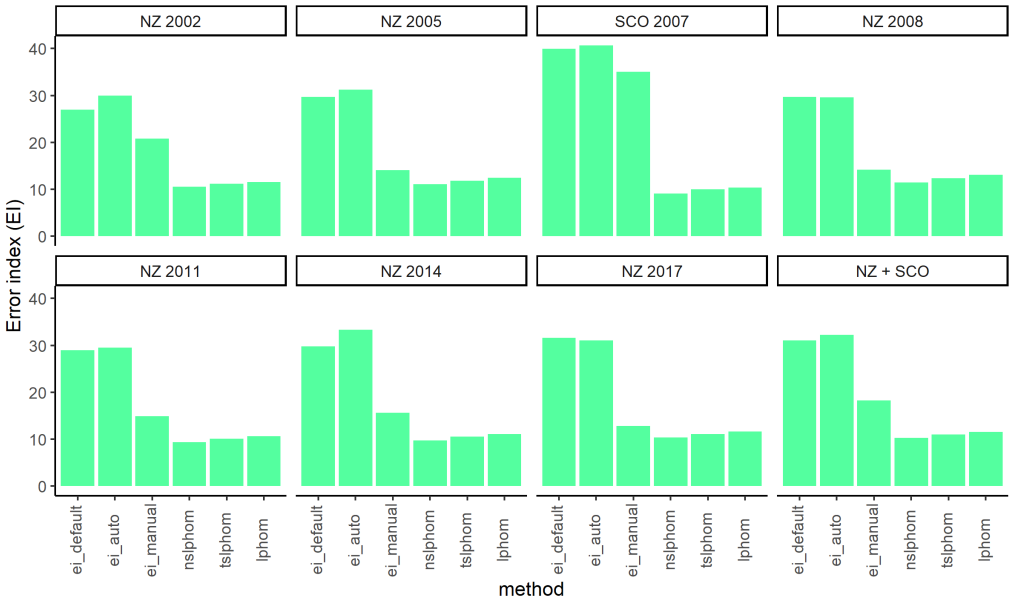

Figure 5. *Graphical representation of average values of EI error measures grouped by election and algorithm in the scenarios attained after merging in Others the election options not surpassing 20% of the vote and randomly merging polling units as described in Section 3. Individual solutions have been attained with the function* `ei.MD.bayes` *of the R package* `eiPack` *(Lau et al., 2020) using three different specifications and the functions* `lphom`*,* `tslphom` *and* `nslphom` *of the R package* `lphom` *(Pavía and Romero, 2021) with default options. Details of the specifications used when applying* `ei.MD.bayes` *can be consulted at the bottom of Table 3.*

(model statistical approach and mathematical programming), the most accurate solutions in our databases. The behavior of both sets of solutions, however, is not homogeneous, presenting important variations among datasets within and between algorithms. In fact, as can be seen in Figure 6, which displays graphically a summary of the average values of EI and WPE in each database for the `ei_manual` and `nslphom` solutions, although `ei_manual` and `nslphom` present (on average) predictions of equivalent quality when the number of available units is large enough, both start to differ clearly when the amount of available data decreases, with the `ei_manual` solutions deteriorating faster. In this section, we look at the analysis in more detail. Focusing exclusively on these two procedures, we investigate, on the one hand, the factors that influence their global accuracies and their differences in accuracy and, on the other hand, the characteristics of the estimates obtained by both algorithms for the fractions p_{jk}. The insights extracted from these latter analyses might open a way forward for exploring how to improve a forecast by combining solutions.

Specifically, after analyzing the distributions of EI and WPE values obtained by both procedures in the entire set of datasets, we investigate the relationship between the accuracies obtained and some of the main characteristics associated with each dataset. With this, we aim to determine what the relative impact of each feature is and to understand

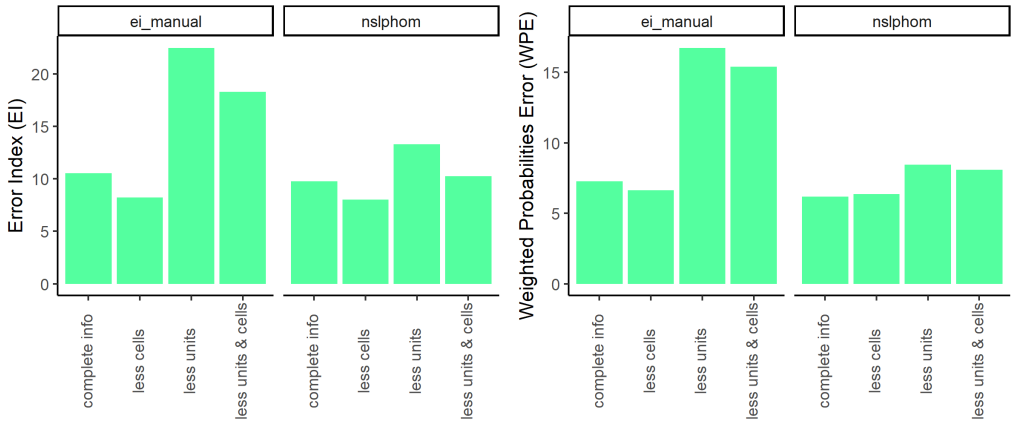

Figure 6. *Graphical representation of global average values of EI and WPE error measures grouped by database for* `ei_manual` *and* `nslphom`*. Individual solutions have been attained with the function* `nslphom` *of the R package* `lphom` *(Pavía and Romero, 2021) with default options and the function* `ei.MD.bayes` *of the R package eiPack (Lau et al., 2020) with customized options. Details of the specification used when applying* `ei.MD.bayes` *can be consulted at the bottom of Table 3.*

under what circumstances each of the methods could work better. This study, focused on the analysis of global accuracies, is complemented by a more detailed look at the cells of the matrices. The second subsection of this section is dedicated to analyzing the quality and properties of the estimates of the fractions p_{jk} that are obtained with both procedures. The analysis is relevant because, according to some authors (e.g., Upton, 1978; Johnston and Hay, 1983), the methods based on mathematical programming have a tendency to predict extreme fractions; the opposite bias attributed by Romero and Pavía (2021) to `ei.MD.bayes`. In the last subsection, we take advantage of these insights to propose a simple rule that can be used to improve forecasts in certain circumstances.

6.1. Factors impacting on the accuracy of the procedures

Figure 6 suggests the existence of important differences in terms of accuracy in the solutions generated by `ei_manual` and `nslphom` and that these depend on the characteristics of the electoral processes under study. Figure 7, where the distributions obtained for *EI* and *WPE* with both procedures are plotted in the 1972 datasets analyzed, clearly shows the existing variability in the solutions reached by each method and between methods (in Table S4 of the supplementary material the interested reader can consult a statistical summary of both distributions). For example, focusing on *EI* (the conclusions for *WPE* would be very similar, see Table S4), we observe that the errors associated with `nslphom` are, on average, more than 4 points lower than those of `ei_manual`. Another interesting observation is that `ei_manual` errors are significantly more dispersed than those obtained by `nslphom`, with respective standard deviations of 10.8 and 4.6. Both results confirm a fact already discussed above: `nslphom` is in this database not

only somewhat better on average but it is also more robust. In fact, although the distance between the medians is much lower than that observed for the means, just 0.76, it continues to be statistically significant, with a p-value smaller than 0.000001 in the sign test for paired data.

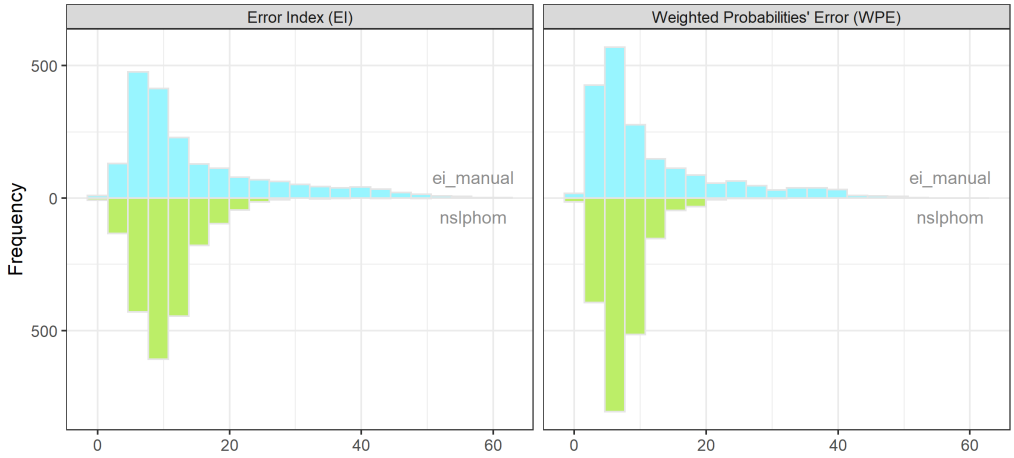

Figure 7. *Histograms of the distributions of the error measures (EI left panel and WPE right panel) linked to the solutions attained after running* `nslphom` *with default options and* `ei.MD.bayes` *with the* `ei_manual` *specification (see the bottom of Table 3 for details) in the 1972 datasets analysed in this research (see Section 3 for details).*

Figure 7 (and Table S4) clearly show that there is a high variability in the accuracies of the results obtained, so it is worth asking what the factors are that would explain, at least in part, the high variability observed within and between methods. Using multiple regression models with EI and WPE as response variables, in this subsection we study the impact that some of the main variables that characterize the scenarios considered have on accuracy. Given the great diversity we have (part of which can be seen in Table 2), we consider this analysis will give us general results regarding the behavior of the two methods rather than about idiosyncrasies of the particular data analyzed, although this cannot be completely discarded.

In addition to variables already considered throughout this paper related to the amount of information available, I, or the complexity associated with the problem, JK, other factors, such as the variability or the degree of dependence presented by the data, have also been proposed in the literature as determinants of the quality of the estimates. Table 4 details the variables considered. Table S5 in the supplementary material presents a statistical summary of the values obtained for the nine variables introduced in Table 4 in the 1972 datasets analyzed, and Table S6 offers the corresponding correlation matrix. A high correlation (0.86) can be seen between both measures of across unit variances on the patterns of votes, *var.Part* and *var.Cand*, with the correlation between *std.Part* and *std.Cand* also being high (0.62). In any case, given the large sample size, we do not expect this to pose a problem in interpreting the models obtained.

Table 4. *Features considered in the models.*

Variable	Description
I	Number of units. Indicator of the quantity of information.
JK	$J \times K$, number of cells in the matrix. Indicator of the complexity of the problem.
JKratio	Quotient J/K. This captures the impact of the asymmetric role played by the two dimensions of the transfer matrix. The algorithms estimate the parameters of J (multinomial) distributions, each one of dimension $K-1$.
HET	Actual heterogeneity index. This measures the degree of non-compliance of the homogeneity hypothesis: $HET = 50(\sum_{ki} \lvert \sum_j N_{j\cdot i} p_{jk} - N_{\cdot ki} \rvert / \sum_{ij} N_{j\cdot i}$. Although this coefficient cannot be computed in regular applications (as the transfer matrix is unknown), it may be estimated.
$Chi2$	Standardized χ^2-Pearson statistic of independence of the global matrix of counts. This measures the degree of dependence between the row and column categories: $Chi2 = \sum_{jk}(N_{jk\cdot} - N_{\cdot k\cdot}N_{j\cdot\cdot})^2/[(J-1)(K-1)\sum_{jk}(N_{\cdot k\cdot}N_{j\cdot\cdot})]$. Although this coefficient cannot be computed in regular applications, it may be estimated.
var.Part	Compositional total variance (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015) of the marginal row distributions in the I units. This measures to what extent party vote supports are different across units: $(2J)^{-1}\sum_{j,j'}^{J} Var(\{log(N_{j\cdot i}/(N_{j'\cdot i})\}_i)$.
var.Cand	Compositional total variance of the marginal column distributions in the I units. This measures to what extent candidacies vote supports are different across units: $(2J)^{-1}\sum_{k,k'}^{K} Var(\{log(N_{\cdot ki}/(N_{\cdot k'i})\}_i)$.
std.Part	Standard deviation of the distribution of percentages of votes to parties in the whole electoral space. Indicator of the degree of vote concentration/variability among parties: $sd(\{N_{j\cdot\cdot}/N_{\cdots}\}_j)$
std.Cand	Standard deviation of the distribution of percentages of votes to candidacies in the whole electoral space. Indicator of the degree of vote concentration/variability among candidacies: $sd(\{N_{\cdot k\cdot}/N_{\cdots}\}_k)$

Source: compiled by the authors.

In order to facilitate the interpretation of the parameters of the fitted models, all the explanatory variables have been standardized, to zero mean and standard deviation 1. In this way, the relative importance of each variable can be directly assessed as it is proportional to the value estimated for its coefficient in the regression model. Approximately, this value multiplied by four quantifies the expected variation in the response variable due to the fluctuation in the sample of the variable considered. Table 5 shows the coefficients of the fitted models. Tables S7 to S12 in the supplementary material show the obtained models in more detail.

Table 5. *Impact of different electoral features on ecological inference solutions' accuracy.*

	Response variable: *EI*			Response variable: *WPE*		
Variable	`nslphom`	`ei_manual`	difference	`nslphom`	`ei_manual`	difference
Constant	10.2110^{***}	14.5998^{***}	4.3888^{***}	7.2631^{***}	11.3220^{***}	4.0589^{***}
I	-1.2926^{***}	-1.9574^{***}	-0.6648^{**}	-0.9977^{***}	-1.5012^{***}	-0.5035^{**}
JK	1.5147^{***}	2.9153^{***}	1.4006^{**}	-0.6129^{**}	-0.0235	0.5894
*JK*ratio	0.7677^{**}	0.6548	-0.1129	1.5018^{***}	2.0344^{***}	0.5326
HET	2.4321^{***}	0.9372^{***}	-1.4949^{***}	1.7783^{***}	0.1746	-1.6037^{***}
*Chi*2	-0.7034^{***}	-1.2736^{***}	-0.5702^{*}	-0.1495	-0.6517^{**}	-0.5022^{*}
var.Part	0.4160^{*}	-3.0448^{***}	-2.6288^{***}	-0.2221	-2.6846^{***}	-2.4625^{***}
var.Cand	-1.8088^{***}	-1.1346^{**}	0.6742	-1.2098^{***}	-0.2001	1.0097^{**}
std.Part	0.4046^{***}	-1.8130^{***}	-2.2176^{***}	0.2740^{***}	-1.8953^{***}	-2.1693^{***}
std.Cand	-0.6001^{***}	-2.5661^{***}	-1.9660^{***}	-0.3760^{***}	-2.1921^{***}	-1.8160^{***}
Adjusted $R^2(\%)$	42.48	30.97	21.44	28.76	26.52	21.65
Std resid. error	3.49	8.97	9.29	2.89	8.16	8.43

Source: compiled by the authors. All the predictor variables were standardized before fitting the models to make comparisons of coefficients easier. ***, p-value < 0.01; **, p-value < 0.05; *, p-value < 0.10. More details of the fitted models can be consulted in Tables S7 to S12 in the supplementary material.

A total of six models were adjusted in order to identify the variables that impact on the quality of predictions (see Table 5). For each discrepancy measure (*EI* and *WPE*) and also for their differences, we adjusted a model to the errors obtained with each of the algorithms (`nslphom` and `ei_manual`). We now focus on analyzing the results obtained for the models using *EI* as the response variable, since the interpretations with *WPE* are similar.

Of the nine variables considered and taking as reference a p-value smaller than 0.01, seven would be selected when analyzing the errors that `nslphom` makes (see the first column of estimates in Table 5). All variables, except *JKratio* and *var.Part*, show a statistically significant impact (p-value < 0.01). Together these variables explain 42% of the observed variability. The complexity of the problem (*JK*), its heterogeneity (*HET*) and the variability across units of the target marginal distributions (*var.Cand*) are revealed as the variables with the greatest effect. Specifically, as expected, the error grows as the complexity of the problem increases and there is greater heterogeneity. Likewise, the errors decrease when there is greater variability in the marginal target distributions. Along with these variables, the amount of information available (*I*), the standard deviations of the global distributions of parties and candidates (*std.Part* and *std.Cand*) and the degree of dependency (*Chi*2) between parties and candidates are also significant. Of these variables, the amount of information is the one that has the greatest impact, and with the expected sign. The error grows as the amount of information available decreases.

The next column offers the adjusted model when analyzing the errors associated with the predictions obtained with `ei_manual`. On this occasion, the model has less explanatory power. However, the same variables identified in the previous model are maintained, and with the same signs. Using 0.01 as cutoff for significance, the main

change lies in the inclusion of the variable *var.Part*, which measures the variability in the marginal distributions of origin. This result is in line with Wakefield (2004), who also in a Bayesian framework states that having smaller within-area variability among row proportions leads to more accurate estimates of fractions. As a rule, it can be seen that in both models the error grows with the complexity of the problem, when the amount of information available decreases or when there is more heterogeneity (i.e., there is more variability between units in the transfer matrices), while the error decreases when there is a greater variety in the data (variance across units) and when there is a greater relationship between the options of the rows and columns. All these variables had already been identified, in one way or another, as determinants for the quality of the estimates (e.g., King, 1997; Park et al., 2014; Klima et al., 2016; Plescia and De Sio, 2018). The relative importance of each of them, however, varies for both methods. It is worth highlighting the fact that the variable *var.Part* which measures the diversity in the marginal distributions of origin, previously identified as a key in the (Bayesian) ecological inference literature, does not appear as a determinant for `nslphom`, where it is subsumed by the variable *var.Cand* which measures the diversity in the marginal target distributions.

In comparative terms, and focusing now on the analysis of the differences (see third column of estimates in Table 5), we can see that although the impact of the amount of information available (I) and of the variability across units in the target distributions (*var.Cand*) affects both methods in a similar way, other variables such as heterogeneity or complexity of the problem do not. The `nslphom` algorithm is more sensitive to non-fulfilment of the homogeneity hypothesis on which it is based, while, in contrast, the `ei_manual` suffers more when the complexity of the problem increases. Likewise, although both methods depend on the variability between the marginal distributions of the territorial units (note that if *var.Cand* or *var.Part* were null, neither of them would be able to reach a solution), `ei_manual` has a greater dependence on *var.Part*, the variability across units between the row marginal distributions. The rest of the variables also have a greater impact on the quality of the `ei_manual` estimates; their estimates improve relatively when there are more differences in sizes between origin and destination options and a greater degree of dependence between them.

Finally, in order to study the possible non-linearity of the effects of the different variables, we also estimate new models in which we consider, in addition to the variables detailed in Table 4, their squares as predictors. The results of these new models, available in Tables S13 to S15 of the supplementary material, reveal the existence of significant quadratic effects for almost all the variables considered; the signs of the curvatures being contrary to those observed for the corresponding linear effect. The conclusion from this is that the estimated effects on EI of an increase in value of the different explanatory variables are especially acute for low values, but diminish as the values increase.

6.2. An analysis of the errors in the estimation of p_{jk}

Once the global adjustments of the matrix forecasts have been analyzed in depth, we focus on the individual cell estimates. In the reference set of 493 elections, a total of

14158 proportions, p_{jk}, were estimated. The results associated with the datasets obtained by random merging of units and/or election options are not considered in this analysis since the collapses do not modify the actual p_{jk} values. The left and middle panels of Figure 8 show the histograms, real and estimated, for `ei_manual` and `nslphom` of the 14158 p_{jk} coefficients. The histograms are found to be slightly bimodal, with a marked accumulation of frequencies in the low values, a continuous decrease as the value of p_{jk} increases and a slight rebound for the highest values.

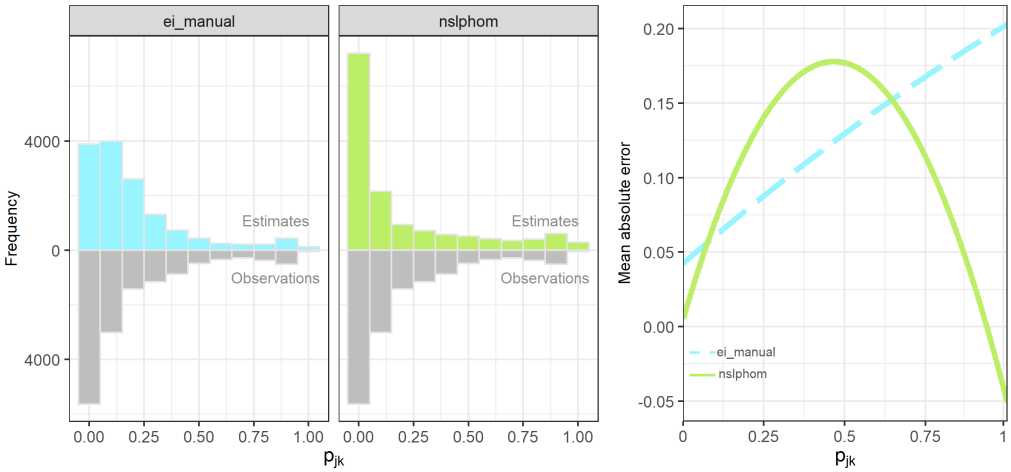

Figure 8. *Histograms of the distributions of* `ei_manual` *(left panel) and* `nslphom` *(centre panel) estimates for p_{jk} and stylized relationships between mean absolute errors of estimates and actual values (right panel). To make the comparisons easier, left and centre panels also offer the actual distributions of the p_{jk} proportions. The displayed* `ei_manual` *and* `nslphom` *estimates correspond to the solutions attained after applying* `ei.MD.bayes` *using the* `ei_manual` *specification and* `nslphom` *with default options to the 493 datasets of the reference database. The curve relationships of the right panel have been obtained after fitting the absolute value errors of the forecasts, $|p_{jk} - \hat{p}_{jk}|$, as a quadratic function of the p_{jk} proportions.*

These forms are a logical consequence of the fact that in split-ticket electoral contexts there are usually close links between the column and row options (in our examples, between parties and candidates), which give rise to the presence of values close to 1 in some rows of the probability matrix (see Table 2), chiefly in the party-rows of the leader candidates. A value close to 1 in a row necessarily implies $C-1$ values close to 0 in that same row. As can be seen in the histograms (see Figure 8), there are numerous values close to 0 and a smaller but relevant number of values relatively close to 1, with still a relevant presence of intermediate values. Intermediate values tend to be more abundant, however, in demographic voting.

The left and center panels of Figure 8 show that the biases attributed in the literature (Upton, 1978; Johnston and Hay, 1983; Romero and Pavía, 2021) to methods based on mathematical programming and `ei.MD.bayes` are manifested in our application: `nslphom` tends to estimate a higher percentage of extreme values and `ei.MD.bayes` to underestimate them. This fact is also reflected in a bias analysis. Table 6 shows the

mean values of the errors of both procedures in the estimation of the p_{jk}, differentiated according to the fact that whether the real values are less than 0.20, greater than 0.80 or intermediate between both limits. On average, the biases are significantly higher for the `ei_manual` than for `nslphom` (see third and fourth columns of Table 6), with a different behavior in both procedures. While `nslphom` tends to overestimate high p_{jk} values and underestimate low values, `ei_manual` tends to overestimate low values and underestimate high values.

Table 6. *Average biases and mean absolute errors (MAE) grouped by intervals of p_{jk}.*

Range	Number of observations	Average bias (×100)		Average MAE (×100)	
		`nslphom`	`ei_manual`	`nslphom`	`ei_manual`
$0.0 \leq p_{jk} < 0.2$	9407	−0.46	4.04	4.27	5.42
$0.2 \leq p_{jk} < 0.8$	3973	0.64	−6.85	15.27	11.29
$0.8 \leq p_{jk} \leq 1.0$	778	2.37	−13.90	4.29	17.78

Source: compiled by the authors.

The problem with calculating mean biases is that they do not reflect the true magnitude of the errors, as they include individual biases with opposite signs in their calculation. To correct this issue, the last two columns of Table 6 provide the mean values of the errors in absolute values. From these data it is clear that `nslphom` is somewhat more precise than `ei_manual` when estimating low values of p_{jk}, less precise when estimating intermediate values and, above all, much more precise when estimating high values. This is clearly a consequence of their default underlying algorithms: `nslphom` takes as seed the `lphom` solution which tends to favor extreme points of the convex hull of the region of feasible solutions defined by the constraints, whereas `ei.MD.bayes` starts, at the very bottom level of the hierarchy, by stating a symmetric distribution that assumes no prior differences between the fractions in each row. To more clearly visualize the situation, the absolute errors obtained by both procedures are adjusted as a function of p_{jk} using a quadratic model. The results of the adjustments are given in the right panel of Figure 8, with their equations available in Tables S16 and S17 of the supplementary material.

Figure 8 (right panel) shows that, as a rule, the estimation errors of `nslphom` are lower than those of `ei_manual` for values of p_{jk} which are lower than 0.10 and higher than 0.65. However, their errors are higher, on average, for intermediate values. The average superiority of `nslphom` over `ei_manual` in the analyzed examples is partially supported, therefore, by the fact that extreme values tend to be frequent in electoral studies of vote transfer. At the cost of automation, therefore, the analyst could reduce the expected bias committed by `ei.MD.bayes` using priors that place higher probabilities on larger fractions for the cells corresponding to intersections of options naturally related among the row and column categories, such as the party and the candidate of the party in ticket-splitting analysis or the same party in voter transition problems. In a mirror fashion, the analyst could also reduce the expected bias committed by `nslphom` in

intermediate fractions by adding new constraints in the model for them. Constraints that reduce their initial space of feasible values from the whole [0,1] interval to some meaningful subinterval. The latter may be considered as reasonable in demographic voting studies.

6.3. Can estimates be improved by combining nslphom and ei_manual solutions?

The previous analyses give clues as to when `ei_manual` and `nslphom` will generate good solutions and also demonstrate that both methods show complementary biases in the estimates of the p_{jk}. This knowledge could be used to improve, on average, the predictions obtained using either of the two methods separately.

On the one hand, we now know that the solutions generated by `ei.MD.bayes` without customizing priors are, as a rule, not reliable when the number of observations is very low. On the other hand, the results suggest that `nslphom` generates robust solutions in a variety of situations. Both results would lead us to clearly recommend `nslphom` when the number of units for which information is available is low and, in general, when it is difficult to achieve convergence in the MCMC chains on which `ei.MD.bayes` is based.

In the above analyses, we have also learned the effect different characteristics of the analyzed scenario have on the aggregated errors, and have also verified that the errors and biases committed by `ei_manual` and `nslphom` are complementary. This last insight could be used to improve, combining both solutions, the individual predictions obtained by each method. We consider that the solutions of `nslphom` could always enter the equation and that the solutions of `ei.MD.bayes` should not enter if we cannot guarantee convergence in the MCMC chains associated with their solutions. Table 7 offers the result of combining (with the same weights) the solutions achieved with `ei_manual` and `nslphom` in the reference database. As can be seen, the combined solutions are, on average, more accurate than the individual solutions. The exception is the solutions that are achieved for Scotland, where the combined solutions are worse than those obtained with `nslphom`.

A detailed analysis of the solutions achieved for Scotland reveals, as shown in Figure S8 of the supplementary material, that the distribution of errors for the solutions achieved with `ei_manual` presents two populations. This is because the algorithm included in `ei.MD.bayes` only achieves, with the `ei_manual` specification, convergence in about half of the elections. In these scenarios, when `ei.MD.bayes` does not reach convergence, the analyst must decide between two alternatives: consider only the `nslphom` solution or manually tune `ei.MD.bayes` in each of the elections until the convergence of the chains can be guaranteed. This second alternative plays against automation and is quite time-consuming, being almost prohibitive when the number of elections to analyze is very high.

The Scottish results therefore raise an important question about when we can combine the solutions of `ei.MD.bayes` and `nslphom`. The obvious answer would be:

Table 7. *Summary of the performance of the solutions attained in the reference database by averaging* `nslphom` *and* `ei_manual` *solutions.*

Country Year	NZ 2002	NZ 2005	SCO 2007	NZ 2008	NZ 2011	NZ 2014	NZ 2017	NZ + SCO
# of Elections	N = 69	N = 69	N = 73	N = 70	N = 70	N = 71	N = 71	N = 493
Avg. # of units	$\bar{I}$ = 83.2	$\bar{I}$= 81.8	$\bar{I}$= 70.2	$\bar{I}$= 84.1	$\bar{I}$= 85.7	$\bar{I}$= 81.2	$\bar{I}$= 101.9	$\bar{I}$= 84.0
Avg. # of cells	$\overline{RC}$= 39.5	$\overline{RC}$= 23.8	$\overline{RC}$= 35.2	$\overline{RC}$ = 23.4	$\overline{RC}$= 26.2	$\overline{RC}$= 27.9	$\overline{RC}$= 24.8	$\overline{RC}$= 28.7
Average of *EI* mesasures								
ei_manual	10.75	8.53	23.09	8.34	7.68	7.88	6.93	10.52
nslphom	12.79	9.68	8.86	9.11	9.46	9.69	8.91	9.77
combined	9.39	7.90	14.09	7.44	7.12	7.05	6.87	8.58
Average of *WPE* mesasures								
ei_manual	6.30	5.61	18.47	5.86	4.88	4.86	4.54	7.28
nslphom	7.90	6.09	4.80	6.09	6.26	6.55	5.67	6.18
combined	5.76	5.23	10.22	5.16	4.71	4.49	4.50	5.75

Source: compiled by the authors after applying the function `nslphom` of the R package `lphom` (Pavía and Romero, 2021) with default options and the function `ei.MD.bayes` of the R package `eiPack` (Lau et al., 2020) with arguments `sample = 1000`, `thin = 100`, `burnin = 100000` and the output of function `tuneMD` with `ntunes = 10` and `totaldraws = 100000` as `tune.list` argument to the official data from the New Zealand electoral commission and the Scotland Electoral Office described in Section 3. Combined solutions have been obtained as arithmetic means of the `ei_manual` and `nslphom` solutions.

when we have reached convergence with `ei.MD.bayes`. This brings us back to the starting point: we have to check convergence (a process not easily automatable) and, if this is not achieved, we have to continue testing specifications, with their enormous associated labor and computational costs. To break this cycle, it would be interesting to study if there is a way to use the robust `nslphom` solution to determine 'automatically' when the solution reached by `ei.MD.bayes` is reliable.

7. Discussion and concluding remarks

The problem of forecasting the inner-cells counts of a contingency table just knowing its row and column aggregates outlines a relevant problem in many settings, including economics, epidemiology and marketing, being sociology and political science where it has aroused more interest. Social scientists, politicians and the media, among other agents, are very interested in mapping the transitions in preferences of voters between elections and in knowing how different social groups vote. Surveys are sometimes used to answer these questions. However, they are not always available (as in historical or local elections) and, more importantly, they are not especially reliable in estimating the coefficients p_{jk}. Polls present significant weaknesses in terms of both precision and accuracy (see, e.g., Miller, 1972; King, 1997; Klima et al., 2016; Dassonneville and Hooghe, 2017; Plescia and De Sio, 2018; Romero et al., 2020). Hence, a number of algorithms have been suggested in the literature to estimate from observed aggregate data the fractions p_{jk} and p^i_{jk}. Because aggregate data are readily available, the issue is to ascertain the performance of the different algorithms.

Several papers have focused on studying theoretically under which circumstances the forecasts obtained would be reliable and how the basic models can be modified under specific circumstances (see, e.g., Firebaugh, 1978; Gelman et al., 2001; Greiner and Quinn, 2009; Forcina and Pellegrino, 2019). The aim of this paper has been to assess, from an empirical perspective, the accuracy and efficiency, among other issues, of the two more powerful methods currently available for forecasting R×C ecological tables: on the one hand, the ecological Bayesian approach programmed in the `ei.MD.bayes` function of the `eiPack` R-package (Lau et al., 2020) and, on the other hand, the mathematical programming algorithms available in the `lphom` R-package (Pavía and Romero, 2021).

In this study, we have started from a singular database made up of almost 500 elections, where we have the gold standard for comparison: the real p_{jk} values, a quite unusual issue (Pavía, 2022). From this baseline database, we have created new scenarios of analysis to evaluate how the different algorithms behave in either more stressful or simpler situations. The results show that to obtain satisfactory solutions with `ei.MD.bayes` it is absolutely essential to properly tune its arguments. It is necessary to guarantee convergence in the MCMC chains on which the algorithm implemented in `ei.MD.bayes` is based in order to obtain reliable solutions. This requires adequately qualified analysts and is accompanied by significant time costs in terms of workforce and computational skills. In contrast, the `lphom` functions, especially the `nslphom` function, are capable of producing accurate results in seconds with their default options, which also makes it robust to claims of hacking. In any case, when `ei.MD.bayes` is properly tuned and convergence is reached (although, sometimes this is more difficult, such as when the amount of information available is scarce) its solutions tend to be slightly more accurate than those of `nslphom`.

In terms of robustness, it is obtained that while `ei.MD.bayes` solutions are much more sensitive to the different characteristics of the dataset used, `nslphom` generates satisfactory solutions in a significantly greater range of scenarios. The inferences of `ei.MD.bayes` with default priors are very sensitive to the data-unknowns relationship, deteriorating notably when the number of units is reduced and, more intensively, when the proportion of rows with extreme fractions grows, while `nslphom` is more robust, being quite insensitive to a decrease in the amount of available data.

The fact that `ei.MD.bayes` malfunctions with few units without proper customization and that `nslphom` generates satisfactory solutions even under those circumstances makes `lphom`-based approaches also preferable in terms of data wrangling. In fact, the costs of obtaining and pre-processing data are generally very relevant in actual ecological inference applications and they grow with the number of units. The `ei.MD.bayes` function also requires that $\sum_k N_{j.i} = \sum_j N_{.ki}$ be verified for all units, $\forall i$, which does not always occur naturally, it being necessary therefore to apply some data pre-processing strategies to guarantee the equalities (Klima et al., 2016). The functions in `lphom`, on the other hand, are capable of handling various scenarios with discrepancies in the previous accounting equalities (Pavía, 2023).

In view of all the previous considerations, our recommendation would be to use `nslphom` as a reference algorithm and to also use `ei.MD.bayes` when we are able to guarantee the convergence of the MCMC chains in the solution provided. In this case, it would even be a good idea to combine both solutions since the biases committed by both functions in the estimation of the coefficients p_{jk} are complementary. While `nslphom` tends to overestimate high p_{jk} values and underestimate low values, `ei_manual` tends to overestimate low values and underestimate high values. This result prompts us to tackle a new line of research to find ways to determine the weights with which the solutions of both functions should be combined to obtain more accurate joint solutions.

We have seen that the accuracy of the solutions achieved by both procedures depends on a set of variables that can be calculated a priori, from the observed data. For example, the `nslphom` algorithm is more sensitive to non-compliance with the homogeneity hypothesis, while `ei_manual` suffers more when the number of units decreases or when the complexity of the problem increases. It would be interesting to study if this insight could be used, when the convergence of the MCMC chains is guaranteed, to determine an optimal weight structure that maximizes the quality (accuracy) of an estimate based on a weighted mean.

Considering the previous idea further, and taking into account that, on the one hand, one of the main weaknesses of the approach implemented in `ei.MD.bayes` lies in the fact that its arguments need to be correctly tuned and, on the other hand, that `nslphom` usually produces reasonable solutions, although slightly worse than `ei.MD.bayes` solutions when this is properly tuned and converges, another line of research worth exploring would be to study whether the use of `ei.MD.bayes` could be automated by defining the priors of its Bayesian specification using the solution reached with `nslphom`. The outputs of `nslphom` could be employed to generate (overdispersed) priors for the `ei.MD.bayes` hyperparameters, including the possibility of using them to produce proper starting values for the α_{jk} and p^i_{jk}, which can be declared to `ei.MD.bayes` through its `start.alphas` and `start.betas` arguments.

The idea would be to study if this strategy would allow better solutions to be reached combining the strengths of both approaches in another way without paying the price of automation. Another advantage of this approach would be that it allows a more natural way of measuring the uncertainty of the estimates. Measures of uncertainty always relevant, that in some contexts, such as in US voting rights litigation, are extremely important. This approach, however, will not come without drawbacks. Using the `nslphom` output to define the `ei.MD.bayes` priors would not produce an authentic Bayesian estimate, since in this scenario the priors to be used by `ei.MD.bayes` would have been generated from the same data that it is going to employ to update them. In this case, this two-step strategy could be exclusively observed as an optimization method, but not as a proper Bayesian approach. Even though, using `nslphom` output to generate starting values for the MCMC chains does make sense, since it should lead to more efficient convergence and better tuning parameters.

In our discussion we have placed certain emphasis on automation (after all, we are dealing with a large number of elections) which is particularly relevant, for instance, in election night analysis. Nevertheless, depending on the context and the ultimate use of the estimates, making inferences beyond filling in the unobserved inner cells of the tables can be more than necessary (for example, in voting rights litigation or in academic studies), and this is more easily accomplished using a full statistical model than a mathematical programming algorithm. Because aggregation involves the loss of information at the individual level, any single approach to ecological inference requires some assumptions, with the success of the effort partially depending on these. Hence, in our view, it pays for the analyst to have a variety of methods that can be used depending on the purpose of the analysis and the logistic, human-resources and time constraints, and also for exploring the data. When different models lead to qualitatively similar conclusions, one can consider the results robust to the different sets of assumptions. But, when various models yield different conclusions, the analyst should, conditional on the ultimate aim of the estimates and/or the circumstances, examine the impact of the different assumptions on the conclusions or make her/his decisions with the aid of this and other comparative studies.

Acknowledgments

The authors wish to thank Carolina Plescia for providing us with the electoral Scottish data handled in this paper and two anonymous reviewers and the editors for their really valuable comments and suggestions. We are grateful to M. Hodkinson for translating and revising the English of the paper and Priscila Espinosa for her tips about LaTeX. This research has been supported by Consellería d'Innovació, Universitats, Ciència i Societat Digital, Generalitat Valenciana [grant number AICO/2021/257] and by the Ministerio de Economía e Innovación [grant number PID2021-128228NB-I00].

Availability of data and material

The New Zealand data used in this research is publicly available on the website http://www.electionresults.org.nz. The Scottish data handled in this paper was provided by Carolina Plescia via personal communication. See also https://links.uv.es/72uQiop, DOI: 10.17605/OSF.IO/DY2SE.

Code availability

Using as a base some of the functions included in the R-packages `eiPack` (version 0.2-1) and `lphom` (version 0.1.3), the ad-hoc R-code employed to apply the assessed algorithms to the particular data analysed in this research is available, with comments in Spanish, in the html files available in https://links.uv.es/Htm570y, DOI: 10.17605/OSF.IO/ZAQH3.

References

Allport, F. H. (1924). The group fallacy in relation to social science. *American Journal of Sociology* 29(6), 688–706.

Barreto, M., Collingwood, L., Garcia-Rios, S., and Oskooii, K. A. R. (2022). Estimating candidate support in Voting Rights Act cases: Comparing iterative EI and EI-RC methods. *Sociological Methods & Research* 51, 271–304.

Brown, P. J., and Payne, C. D. (1986). Aggregate data, ecological regression, and voting transitions. *Journal of the American Statistical Association* 81(394), 452–460.

Choirat, C., Gandrud, C., Honaker, J., Imai, K., King, G., and Lau, O. (2017). *Zelig: Everyone's Statistical Software, Version 5.0-15*. URL: http://ZeligProject.org

Collingwood, L., Oskooii, K., Garcia-Rios, S., and Barreto, M. (2016). eiCompare: Comparing ecological inference estimates across EI and EI:RxC *The R Journal* 8, 92–101.

Dassonneville, R., and Hooghe, M. (2017). The noise of the vote recall question: The validity of the vote recall question in panel studies in Belgium, Germany, and the Netherlands. *International Journal of Public Opinion Research* 29(2), 316–338.

Ferree, K. E. (2004). Iterative approaches to RxC ecological inference problems: where they can go wrong and one quick fix. *Political Analysis* 12(2), 143–159.

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review* 43, 557–572.

Forcina, A., and Pellegrino, D. (2019). Estimation of voter transitions and the ecological fallacy. *Quality & Quantity* 53, 1859–1874.

Gelman, A., Park, D.K., Ansolabehere, S., Price, L. C., and Minnite, L. C. (2001). Models, assumptions and model checking in ecological regression. *Journal of the Royal Statistical Society, Series A* 164(1), 101–118.

Gehlke, C. E., and Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association* 29(185A), 169–170.

Goodman, L. A. (1953). Ecological regressions and the behaviour of individuals. *American Sociological Review* 18, 663–664.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology* 64(6), 610–625.

Gosnell, H. F., and Gill, N. N. (1935). An Analysis of the 1932 Presidential vote in Chicago. *The American Political Science Review* 29, 967–984.

Greiner, D. J., and Quinn, K. M. (2009). RxC ecological inference: Bounds, correlations, flexibility, and transparency of assumptions. *Journal of the Royal Statistical Society, Series A* 172(1), 67–81.

Greiner, D. J., and Quinn, K. M. (2010). Exit polling and racial bloc voting: Combining individual-level and RxC ecological data. *The Annals of Applied Statistics* 4, 1774–1796.

Imai, K., King, G., and Lau, O. (2008). Toward a common framework for statistical analysis and development. *Journal of Computational and Graphical Statistics* 17, 892–913.

Johnston, R. J., and Hay, A. M. (1983). Voter transition probability estimates: An entropy-maximizing approach. *European Journal of Political Research* 11, 93–98.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.

King, G., Rosen, O., and Tanner, M. A. (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research* 28, 61–90.

Klein, J. M. (2019). *Estimation of Voter Transitions in Multi-Party Systems. Quality of Credible Intervals in (hybrid) Multinomial-Dirichlet Models.* Master Thesis Dissertation. Ludwig-Maximilians-Universität München.

Klima, A., Thurner, P. W., Molnar, C., Schlesinger, T., and Küchenhoff, H. (2016). Estimation of voter transitions based on ecological inference: an empirical assessment of different approaches. *AStA - Advances in Statistical Analysis* 100, 133–159.

Klima, A., Schlesinger, T., Thurner, P. W., and Küchenhoff, H. (2019). Combining aggregate data and exit polls for the estimation of voter transitions. *Sociological Methods & Research* 48, 296–325.

Lau, O., Moore, O. R. T., and Kellermann, M. (2007). eiPack: RxC ecological inference and higher-dimension data management. *The R Journal* 7, 43–47.

Lau, O., Moore, O. R. T., and Kellermann, M. (2020). *eiPack: Ecological Inference and Higher-Dimension Data Management. R package version 0.2-1*. https://CRAN.R-project.org/package=eiPack

Manski, C. F. (2007). *Identification for Prediction and Decision*. Harvard University Press.

Martín, J. (2020). *Análisis de la incertidumbre en la estimación de la movilidad electoral mediante el procedimiento lphom*. PhD Dissertation. Universidad Politécnica de Valencia.

Miller, W. L. (1972). Measures of electoral change using aggregate data. *Journal of the Royal Statistical Society, Series A* 135, 122–142.

Ogburn, W. F., and Goltra, I. (1919). How women vote. *Political Science Quarterly* 34, 413–433.

Park, W., Hanmer, M. J., and Biggers, D. R. (2014). Ecological inference under unfavorable conditions: straight and split-ticket voting in diverse settings and small samples. *Electoral Studies* 36, 192–203.

Pavía, J. M. (2022). ei.Datasets: Real datasets for assessing ecological inference algorithms. *Social Science Computer Review* 40, 247–260.

Pavía, J. M. (2023). Adjustment of initial estimates of voter transition probabilities to guarantee consistency and completeness. *SN Social Sciences*, 3, 75.

Pavía, J. M., and Aybar, C. (2020). Electoral mobility in the 2019 elections in the Valencian region. *Debats. Journal on Culture, Power and Society*, 134, 27–51.

Pavía, J. M., and Romero, R. (2021). *lphom: Ecological Inference by Linear Programming under Homogeneity. R package version 0.1.3*. https://CRAN.R-project.org/package=lphom

Pavía, J. M., and Romero, R. (2022). Improving estimates accuracy of voter transitions. Two new algorithms for ecological inference based on linear programming. *Sociological Methods & Research*, online available, https://doi.org/10.1177%2F00491241221092725

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data.* Chichester: John Wiley & Sons, Ltd.

Petropoulos, F., et al. (2022). Forecasting: Theory and practice. *International Journal of Forecasting* 38, 705–871.

Plescia, C., and De Sio, L. (2018). An evaluation of the performance and suitability of RxC methods for ecological inference with known true values. *Quality & Quantity* 52, 669–683.

Robert, C. P., and Casella, G. (2004). *Monte Carlo Statistical Methods.* Springer.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review* 15, 351–357.

Romero, R., Pavía, J. M., Martín, J, and Romero, G. (2020). Assessing uncertainty of voter transitions estimated from aggregated data. Application to the 2017 French presidential election. *Journal of Applied Statistics* 47(13-15), 2711–2736.

Romero, R., and Pavía, J. M. (2021). Estimating vote party entries and exits by ecological inference. Mathematical programming versus Bayesian statistics. *BEIO: Boletín de Estadística e Investigación Operativa* 34, 85–97.

Rosen, O., Jiang, W., King, G., and Tanner, M. A. (2001). Bayesian and frequentist inference for ecological inference: The RxC Case. *Statistica Neerlandica* 55(2), 134–156.

Roy, V. (2020). Convergence diagnostics for Markov chain Monte Carlo. *Annual Review of Statistics and Its Application* 7, 387–412.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society - Series B* 13(2), 238–241.

Thomsen, S. R. (1987). *Danish elections, 1920-79: A Logit Approach to Ecological Analysis and Inference.* Aarhus: Politica.

Upton, G. J. G. (1978). A note on the estimation of voter transition probabilities. *Journal of the Royal Statistical Society, Series A.* 141, 507–512.

Wakefield, J. (2004). Ecological inference for 2x2 tables (with discussion). *Journal of the Royal Statistical Society, Series A.* 167, 385–445.

SORT 47 (1) January-June 2023, 187-204 DOI: 10.57645/20.8080.02.5

Inference on the symmetry point-based optimal cut-off point and associated sensitivity and specificity with application to SARS-CoV-2 antibody data

A.M. Franco-Pereira[1,2], M.C. Pardo[1,2], C.T. Nakas[3,4] and B. Reiser[5]

Abstract

In the presence of a continuous response test/biomarker, it is often necessary to identify a cut-off point value to aid binary classification between diseased and non-diseased subjects. The symmetry-point approach which maximizes simultaneously both types of correct classification is one way to determine an optimal cut-off point. In this article, we study methods for constructing confidence intervals independently for the symmetry point and its corresponding sensitivity, as well as respective joint nonparametric confidence regions. We illustrate using data on the generation of antibodies elicited two weeks post-injection after the second dose of the Pfizer/BioNTech vaccine in adult healthcare workers.

MSC: *62F10, 62G07, 65C05, 62P10.*

Keywords: *Empirical likelihood function, Empirical chi-square function, Box-Cox transformation, Confidence regions, Sensitivity, Specificity.*

1. Introduction

Let X_1 and X_2 denote continuous response variables (biomarkers) for two user-defined groups (e.g. non-diseased versus diseased subjects), and let F_{X_1} and F_{X_2} be the cor-

[1] Department of Statistics and O.R., Complutense University of Madrid.

[2] Instituto de Matemática Interdisciplinar (IMI), Complutense University of Madrid, Plaza de Ciencias 3, 28040-Madrid, Spain.

[3] Laboratory of Biometry, Department of Agriculture, Crop Production and Rural Environment, University of Thessaly, Phytokou street, 38446 Volos, Greece.

[4] University Institute of Clinical Chemistry, Inselspital, Bern University Hospital, University of Bern, Freiburgstrasse, 3010 Bern, Switzerland.

Received: January 2023
Accepted: May 2023

responding probability distribution functions. Using a cut-off point c to decide that a subject is non-diseased when a marker measurement is less than c and that the subject is diseased otherwise, the specificity of the marker is $spec(c) = P(X_1 \leq c)$ and the sensitivity of the marker is $sens(c) = P(X_2 > c)$. We further make the standard assumption that larger values of the marker are more indicative of disease. The Receiver Operating Characteristic (ROC) curve is defined by $sens(c)$ versus $1 - spec(c)$ as c varies over the support of the response variable values. The symmetry point is the point c_s where $spec(c_s) = sens(c_s)$, which is where the ROC curve and the line $sens = 1 - spec$ intersect. The symmetry point approach for cut-off point selection has appeared in a wide range of recent applications in practice (see e.g. Arnone et al. 2020; Le, Ku and Jun, 2021; Sande et al. 2021; Sekgala et al. 2022) owing to its optimality properties rising from game theory considerations (Sanchez, 2017). However, in these studies, the estimated symmetry point is reported without confidence intervals, emphasizing the importance of having efficient and effective methods of statistical inference on the symmetry point cut-off and its sensitivity along with accessible software for implementation.

To our knowledge, the only two proposals for constructing confidence intervals (CIs) of the symmetry point-based optimal cutpoint and its associated sensitivity have been given by López-Ratón et al. (2016). They are based on the generalized pivotal quantity (GPQ) and the empirical likelihood (EL), respectively. The authors recommended the use of EL method when the distributions of healthy and diseased populations are unknown. Later, Adimari and Sinigaglia (2020) proposed a nonparametric method that provides joint confidence regions for the symmetry point-based optimal cutpoint and its associated sensitivity. Their method is also based on EL and uses the fact that the asymptotic distribution of the statistic they use has a chi-squared distribution with two degrees of freedom. We discuss an alternative to these nonparametric methods based on EL as well as parametric approaches for the construction of confidence intervals.

In the following section we present parametric and nonparametric approaches for the construction of confidence intervals for the symmetry point and its associated sensitivity, or equivalently its specificity, separately as well as methods for the construction of simultaneous confidence regions. Simulation studies comparing the different methods are presented in Section 3. In Section 4, an application to data of SARS-CoV-2 antibody levels is presented pertinent to the diagnosis of prior COVID-19 for possibly asymptomatic individuals. We end with a discussion.

2. Methods

2.1. Construction of confidence intervals: parametric approaches

Let $X_{11}, X_{12}, ..., X_{1n_1}$ and $X_{21}, .X_{22}, ..., X_{2n_2}$ denote two random samples of sizes n_1 and n_2 taken from two independent normal distributions with mean μ_i and variance σ_i^2, $i = 1, 2$, respectively. Under this assumption, it follows that the symmetry point c_s satisfies the following equation

$$\Phi\left(\frac{\mu_2-\mu_1+\sigma_1\Phi^{-1}(x)}{\sigma_2}\right)=1-x$$

where $spec(c_s)=1-x$ and Φ is the cumulative distribution function of a variable following a standard normal distribution. After elemental algebra (see López-Ratón et al. (2016)), we obtain the following closed-form expression:

$$c_s=\frac{\mu_1\sigma_2+\mu_2\sigma_1}{\sigma_1+\sigma_2} \tag{1}$$

and

$$spec(c_s)=sens(c_s)=\Phi(\delta_s)=\Phi\left(\frac{\mu_2-\mu_1}{\sigma_1+\sigma_2}\right). \tag{2}$$

Both c_s and $sens(c_s)$ are estimated by substituting for the unknown $\mu_1,\mu_2,\sigma_1,\sigma_2$ in the above formulae their maximum likelihood estimates (MLEs), $\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma}_1,\widehat{\sigma}_2$. Sensitivity and specificity are proportions and thus they are bounded between zero and one. As a result, the normal approximation for the construction of confidence intervals described in the classical approach can be inadequate for small samples and may also result in intervals which exceed the bounds. To obtain a (1-α)% confidence interval for $sens(c_s)$ we apply standard normal asymptotic theory on $\widehat{\delta}_s$ which is not bounded, and use $\Phi\left(\widehat{\delta}_s\pm z_{1-\alpha/2}\sqrt{\widehat{Var}\left(\widehat{\delta}_s\right)}\right)$ where $z_{1-\alpha/2}$ refers to the $1-\alpha/2$ percentile of the standard normal distribution. Since $\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma}_1,\widehat{\sigma}_2$ are all independent, using the delta method, we obtain

$$\widehat{Var}(\widehat{c}_s)\approx\left(\frac{\partial\widehat{c}_s}{\partial\mu_1}\right)^2\widehat{Var}(\widehat{\mu}_1)+\left(\frac{\partial\widehat{c}_s}{\partial\mu_2}\right)^2\widehat{Var}(\widehat{\mu}_2)+\left(\frac{\partial\widehat{c}_s}{\partial\sigma_1}\right)^2\widehat{Var}(\widehat{\sigma}_1)+\left(\frac{\partial\widehat{c}_s}{\partial\sigma_2}\right)^2\widehat{Var}(\widehat{\sigma}_2)$$

where $\frac{\partial\widehat{c}_s}{\partial\mu_1}=\frac{\widehat{\sigma}_2}{\widehat{\sigma}_1+\widehat{\sigma}_2},\frac{\partial\widehat{c}_s}{\partial\mu_2}=\frac{\widehat{\sigma}_1}{\widehat{\sigma}_1+\widehat{\sigma}_2},\frac{\partial\widehat{c}_s}{\partial\sigma_1}=\frac{\widehat{\sigma}_2(\widehat{\mu}_2-\widehat{\mu}_1)}{(\widehat{\sigma}_1+\widehat{\sigma}_2)^2},\frac{\partial\widehat{c}_s}{\partial\sigma_2}=\frac{\widehat{\sigma}_1(\widehat{\mu}_1-\widehat{\mu}_2)}{(\widehat{\sigma}_1+\widehat{\sigma}_2)^2}$ and

$$\widehat{Var}\left(\widehat{\delta}_s\right)\approx\left(\frac{\partial\widehat{\delta}_s}{\partial\mu_1}\right)^2\widehat{Var}(\widehat{\mu}_1)+\left(\frac{\partial\widehat{\delta}_s}{\partial\mu_2}\right)^2\widehat{Var}(\widehat{\mu}_2)+\left(\frac{\partial\widehat{\delta}_s}{\partial\sigma_1}\right)^2\widehat{Var}(\widehat{\sigma}_1)+\left(\frac{\partial\widehat{\delta}_s}{\partial\sigma_2}\right)^2\widehat{Var}(\widehat{\sigma}_2)$$

where $\frac{\partial\widehat{\delta}_s}{\partial\mu_1}=\frac{1}{\widehat{\sigma}_1+\widehat{\sigma}_2},\frac{\partial\widehat{\delta}_s}{\partial\mu_2}=-\frac{1}{\widehat{\sigma}_1+\widehat{\sigma}_2},\frac{\partial\widehat{\delta}_s}{\partial\sigma_1}=-\frac{\widehat{\mu}_2-\widehat{\mu}_1}{(\widehat{\sigma}_1+\widehat{\sigma}_2)^2},\frac{\partial\widehat{\delta}_s}{\partial\sigma_2}=-\frac{\widehat{\mu}_2-\widehat{\mu}_1}{(\widehat{\sigma}_1+\widehat{\sigma}_2)^2}$. The implied confidence interval for c_s is of the form $\widehat{c}_s\pm z_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{c}_s)}$. We refer to this approach as "δ".

The assumption that the biomarkers are normally distributed can be quite restrictive leading to false results when it is significantly violated. A popular way of extending the parametric approach is the use of the Box-Cox transformation (Box and Cox, 1964) which has been previously employed in the ROC framework (e.g. Faraggi and Reiser

(2002); Fluss, Faraggi and Reiser (2005); Molodianovitch, Faraggi and Reiser (2006); Schisterman et al. (2008); Franco-Pereira, Nakas and Pardo (2020); López-Ratón et al. (2016)) and has been shown to perform very well. The Box-Cox transformation is defined by $X_i^{(\lambda)} = \dfrac{X_i^{\lambda}-1}{\lambda}$ for $\lambda \neq 0$ and $X_i^{(0)} = \log(X_i)$ where it is assumed that $X_i^{(\lambda)} \sim N\left(\mu_i^{(\lambda)}, \sigma_i^{(\lambda)}\right)$. The maximum likelihood estimate (MLE) $\widehat{\lambda}$ of the common transformation parameter λ can be obtained by maximizing the profile likelihood function given in Franco-Pereira et al. (2021). We use $\widehat{c}_s^{(BC)}$ and $\widehat{\delta}_s^{(BC)}$ to denote the estimate of c_s and δ_s obtained above, but using the transformed observations. The estimator $\widehat{c}_s^{(BC)}$ needs to be backtransformed to obtain an estimator of the symmetry point on the original scale. In this approach the "added" variation due to estimating the transformation is not taken into account following the rationale as Schisterman et al. (2008). We refer to this approach as "δ-BC".

Another possibility is to consider a bootstrap-based approach in order to obtain an estimate of the variance of $\widehat{c}_s$ and $\widehat{sens}(c_s)$ and thus to compute the $100(1-\alpha)\%$ confidence interval for c_s and $sens(c_s)$ through the following steps:

Algorithm 1

1. Take a sample with replacement from X_1 and X_2.
2. Carry out the Box-Cox transformation by maximizing the profile likelihood given in Franco-Pereira et al. (2021) for each bootstrap sample.
3. For $i = 1,2$, calculate $\widehat{\mu}_i^{(\lambda)}$ and $\widehat{\sigma}_i^{(\lambda)}$, the MLE of μ_i and σ_i, respectively.
4. Calculate $\widehat{c}_s^{(BC)}$ and $\widehat{\delta}_s^{(BC)}$ in (1) and (2) by replacing the μ_i and σ_i with $\widehat{\mu}_i^{(\lambda)}$ and $\widehat{\sigma}_i^{(\lambda)}$.
5. Back-transform $\widehat{c}_s^{(BC)}$ to obtain the current estimate for the symmetric point on the original scale, denoted by $\widehat{c}_s$.
6. Repeat steps 1-5 B times. Then, based on the B values of $\widehat{c}_s$ and $\widehat{\delta}_s^{(BC)}$, $\widehat{c}_{sb}$ and $\widehat{\delta}_{sb}^{(BC)}$, derive the bootstrap estimate $\widehat{Var}_B(\widehat{c}_s)$ and $\widehat{Var}_B\left(\widehat{\delta}_s^{(BC)}\right)$, respectively.
7. Construct the two-sided $100(1-\alpha)\%$ confidence interval of c_s as $\widehat{c}_s \pm z_{1-\alpha/2}\sqrt{\widehat{Var}_B(\widehat{c}_s)}$ and $sens(c_s)$ as $\Phi\left(\widehat{\delta}_s^{(BC)} \pm z_{1-\alpha/2}\sqrt{\widehat{Var}_B\left(\widehat{\delta}_s^{(BC)}\right)}\right)$.

We refer to this approach as "BC-AN". The bootstrap estimates, $\widehat{c}_{sb}$ and $\widehat{\delta}_{sb}^{(BC)}$ can also be used to obtain bootstrap percentile confidence intervals ("BC-PB") as well as the bias corrected bootstrap confidence interval ("BC-bias"). Note again that these bootstrap estimators do take into account the variation due to the estimation of λ.

2.2. Construction of confidence intervals: nonparametric approaches

López-Ratón et al. (2016) constructed confidence intervals for the symmetry point and its associated sensitivity index using a parametric aproach based on the generalized pivotal quantity (GPQ) and a non-parametric approach based on the empirical likelihood (EL). They recommended the use of the EL method when the distributions of healthy and diseased populations are unknown and this is the approach we are going to consider. Details of this procedure are presented therein. The method is based on the empirical likelihood function which is given by

$$l\left(sens\left(c_s\right),c_s\right)=2n_1\left\{\widehat{F}_{X_1}(c_s)\log\frac{\widehat{F}_{X_1}(c_s)}{sens\left(c_s\right)}+\left(1-\widehat{F}_{X_1}(c_s)\right)\log\frac{1-\widehat{F}_{X_1}(c_s)}{1-sens\left(c_s\right)}\right\} \quad (3)$$
$$+2n_2\left\{\widehat{F}_{X_2}(c_s)\log\frac{\widehat{F}_{X_2}(c_s)}{1-sens\left(c_s\right)}+\left(1-\widehat{F}_{X_2}(c_s)\right)\log\frac{1-\widehat{F}_{X_2}(c_s)}{sens\left(c_s\right)}\right\}$$

where $\widehat{F}_{X_i}$ is the Gaussian kernel estimate of the cumulative distribution function F_{X_i}, $i=1,2$, using the same bandwidth given in López-Ratón et al. (2016). We refer to this approach as "EL". However, in many contexts, the chi-square test statistic works better than the likelihood ratio test statistic. Pardo, Lu and Franco-Pereira (2022) compared test statistics based on the empirical likelihood and chi-squared functions for testing monotone and umbrella orderings, and that based on the chi-square function was the most powerful. Pardo and Pardo (2008) found that the chi-square test statistic outperforms the classical loglikelihood test statistic for selecting a model from a sequence of Generalized Linear Models with binary data. Therefore, in this work we consider the procedure described by López-Ratón et al. (2016) substituting the empirical likelihood function with the empirical chi-square function

$$\Lambda\left(sens\left(c_s\right),c_s\right)=\frac{n_1\left(\widehat{F}_{X_1}(c_s)-sens\left(c_s\right)\right)^2}{sens\left(c_s\right)\left(1-sens\left(c_s\right)\right)}+\frac{n_2\left(\widehat{F}_{X_2}(c_s)-\left(1-sens\left(c_s\right)\right)\right)^2}{sens\left(c_s\right)\left(1-sens\left(c_s\right)\right)} \quad (4)$$

in step 2 of their algorithm. A nonparametric estimator of $sens(c_s)$ is obtained which is in turn used to minimize Equation (4) in c_s and consider the minimum found as the nonparametric estimator of c_s. Then, we resample independently from the original pair of samples B times and repeating the above estimation procedure to obtain B bootstrap estimators of $sens(c_s)$ and c_s. Finally, these estimates are used to construct the CIs by the percentile method. We refer to this approach as "ECS".

2.3. Construction of confidence regions

A joint region provides more precise information about the pair of parameters of interest $(sens(c_s),c_s)$ than the marginal confidence intervals do. In our simulation study reported below (Section 4) the parametric methods did not perform satisfactorily for the marginal

CIs obtained for c_s. As a result, we focus, in this section, on constructing nonparametric confidence regions for the symmetry point and its associated sensitivity index. Adimari and Sinigaglia (2020) proposed an approach based on the computation of the empirical distribution function from the data. As a consequence, the set

$$R_\alpha = \left\{ (sens(c_s), c_s) : l(sens(c_s), c_s) \leq \chi^2_{2,\alpha} \right\}$$

with $l(sens(c_s), c_s)$ defined in (3) and $\chi^2_{2,\alpha}$ is such that $P\left(\chi^2_2 \geq \chi^2_{2,\alpha}\right) = \alpha$, is a confidence region with nominal coverage probability $1-\alpha$ for the $(sens(c_s), c_s)$ point.

Since the asymptotic distribution of $\Lambda(sens(c_s), c_s)$ is the same as the asymptotic distribution of $l(sens(c_s), c_s)$ since $l(x,y) = \Lambda(x,y) + o_p(1)$, an alternative confidence region with nominal coverage probability $1-\alpha$ is given by

$$R^*_\alpha = \left\{ (sens(c_s), c_s) : \Lambda(sens(c_s), c_s) \leq \chi^2_{2,\alpha} \right\}$$

with $\Lambda(sens(c_s), c_s)$ defined in (4).

3. Simulation study

3.1. CIs

A large simulation study was conducted in order to compare the approaches for constructing confidence intervals described in Sections 2.1 and 2.2, namely: δ, δ-BC, BC-AN, BC-PB, BC-bias and ECS. We also compare our approaches with the one based on the empirical likelihood (EL) proposed by López-Ratón et al. (2016). We generated data from Normal, PowerNormal$\left(X_1^{-1/3}, X_2^{-1/3}\right)$, LogNormal, Gamma and Mixed models. The parameters used for each of these scenarios are the same as those given in Table 1 of López-Ratón et al. (2016). We used sample sizes: $n_1 = n_2 = 30, 50, 100$ and the unequal sample size scenarios $(20, 30)$, $(50, 100)$ and $(50, 300)$. The number of Monte Carlo replications utilized was N=1000 and B=500 for the bootstrap technique. Performances of each CI approach were assessed by coverage probability (CP) and mean of interval lengths (widths) for c_s and $sens(c_s)$, respectively.

Figures 1 and 3 summarize graphically the observed coverage probabilities and average widths of the CIs for each scenario for both c_s and sensitivity, respectively. In terms of coverage for c_s, except for the case of the mixture of normal distributions, we notice that the nonparametric methods perform substantially better than the parametric ones even in the normal case. Therefore, the width tends to be larger for the two nonparametric approaches. For normal, gamma and mixture models scenarios, both nonparametric methods, ECS and EL, are similar. ECS provides the best results for PowerNormal and EL for LogNormal. If we focus on the parametric approaches, the "BC-AN" approach outperfoms the others except for the normal mixture scenario for which "δ -BC" is the best but the spread of the observed average values is very large in comparison with the

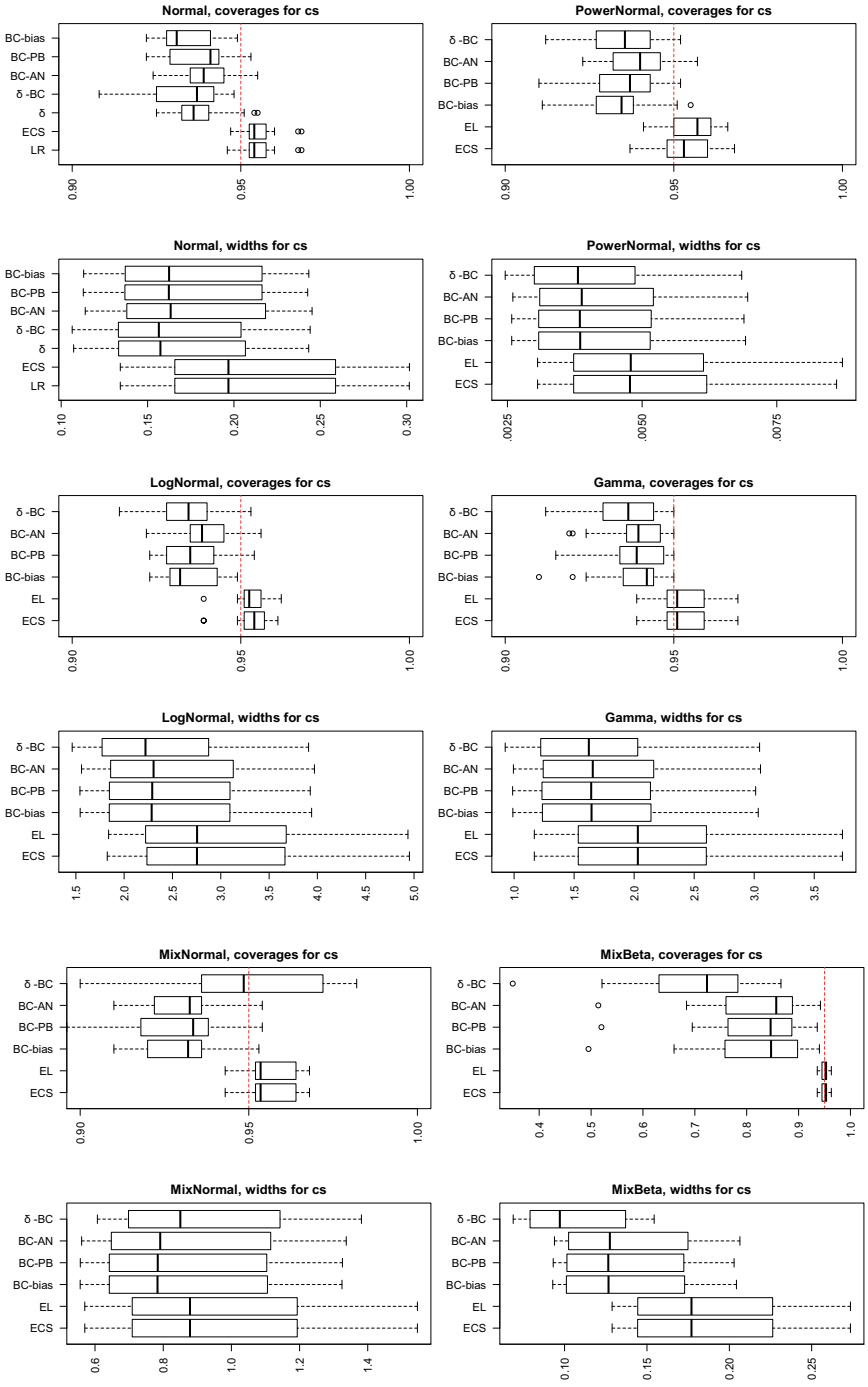

Figure 1. *Boxplots of the coverages and average widths of the confidence intervals for c_s in the different scenarios considered in the simulation study.*

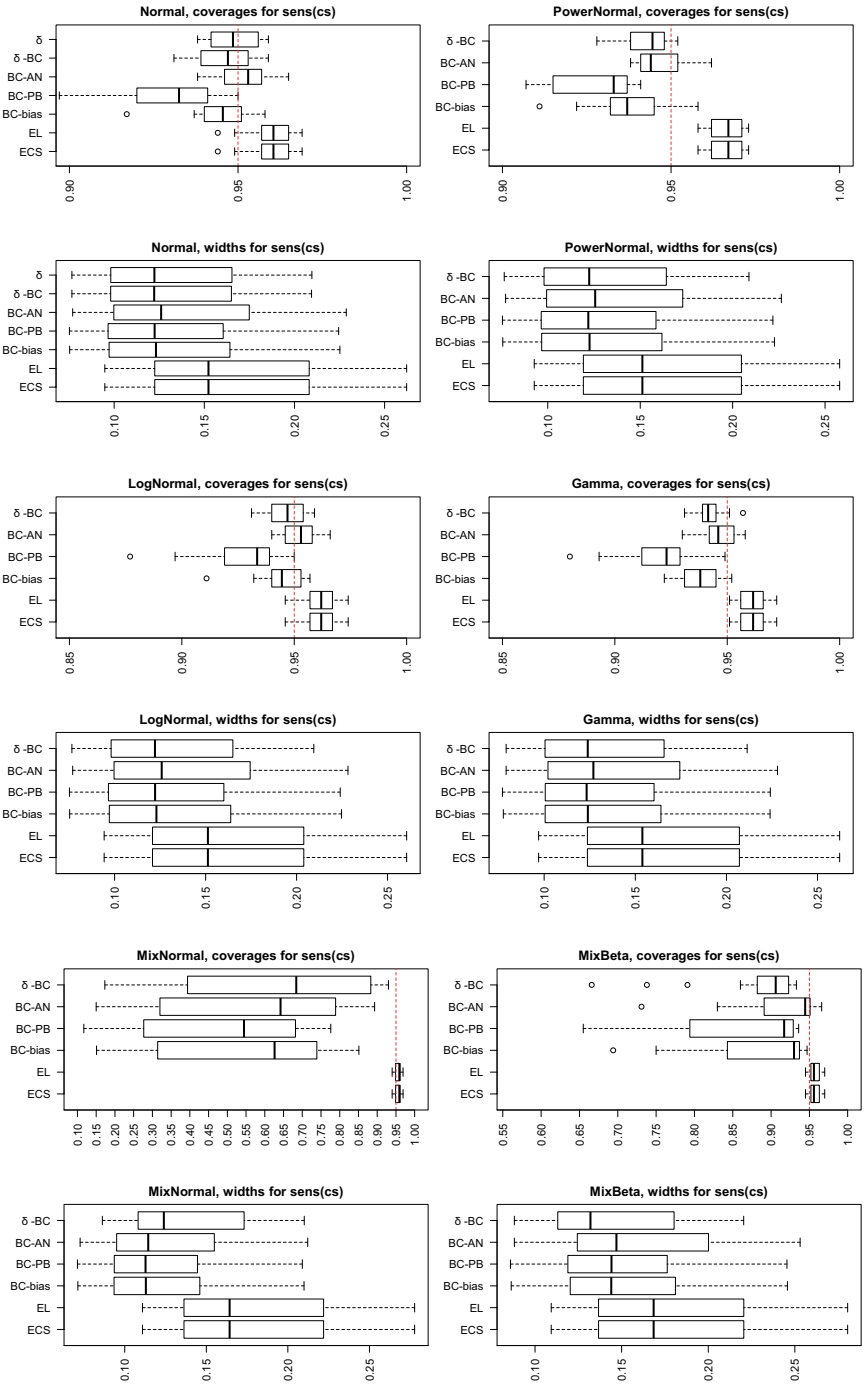

Figure 2. *Boxplots of the coverages and average widths of the confidence intervals for sens(c_s) in the different scenarios considered in the simulation study.*

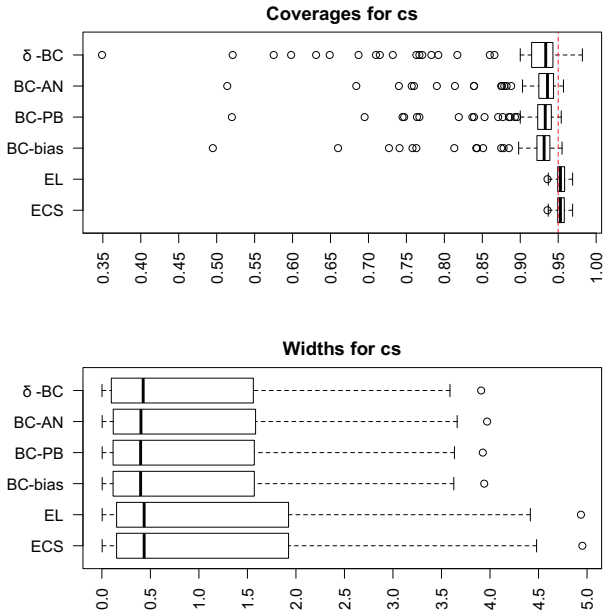

Figure 3. *Boxplot of the coverages and average widths of the confidence intervals for c_s for all scenarios combined.*

others. The parametric approaches tend to be somewhat conservative and their coverages tend to be more variable over the various scenarios and sample sizes sometimes exhibiting quite low coverages while the nonparametric methods have coverages which vary much less. However, looking at the coverages for $sens(c_s)$ except for the normal mixtures, BC-AN seems to do quite well having coverage closer to the nominal and shorter length than the nonparametric methods. In this case, the nonparametric methods tend to be more liberal having more than the nominal coverage which naturally leads to longer average widths.

Therefore, ECS and EL are recommended for constructing CIs for c_s and BC-AN for $sens(c_s)$ as can be seen in Figures 1 through 4, which provide a summary for all the methods merging all the scenarios into box-plots of the observed coverages and CI widths. It is important to take the coverage into account with a head-to-head comparison only being justified for similar coverages. However, the relationship between average length and coverage is not one-to-one, as a result, it is reported throughout for the sake of completeness in the Supplement (Tables 1-24).

3.2. CRs

A second simulation study was conducted to compare the approach based on the empirical likelihood function for constructing confidence regions proposed by Adimari and Sinagaglia (2020) with our proposal based on the chi-squared function given in (4). We generated data from Normal, LogNormal, Beta, Exponential, Gamma and Mixtures of normal distributions. The parameters used for each of these scenarios are the

same as those considered in Adimari and Sinagaglia (2020). We used sample sizes: $n_1 = n_2 = 20, 50, 100$ and the unequal sample size scenarios $(50, 20)$ and $(20, 50)$. The number of Monte Carlo replications was N=10000. The performance of each CR approach was assessed by the proportions of cases falling inside the confidence region, and mean of the confidence regions of the areas. In the simulation study we consider five different scenarios: two scenarios correspond to the normal model, for the third and fourth we use the beta and the gamma models, respectively. Finally, the fifth scenario corresponds to mixture models (see Table 25 in the Supplement).

The results of these simulations, for three levels of nominal coverage $1 - \alpha$, that is, 0.90, 0.95 and 0.99, are shown in Tables 26-41 in the Supplement. Tables 26–33 and Tables 34–41 give the estimated coverage probabilities and estimated areas of the confidence regions, respectively, obtained by using both methods presented in Section 2.3. for each scenario. As expected, the simulated coverage is closer to the nominal when the sample size increases. Our proposal generally provides results with observed coverage closer to the nominal level. However, the observed coverages of the confidence regions based on R^*_α are more variable than R_α for most of the scenarios. Figure 5 provides a graphical presentation of these results as well as a box-plot with the estimated coverage for all scenarios combined. In relation to the areas of the confidence regions, looking at Figure 6, it can be seen that they are very similar for both methods. Therefore, R^*_α approach produces confidence regions with a coverage closer to the nominal level without increasing their area.

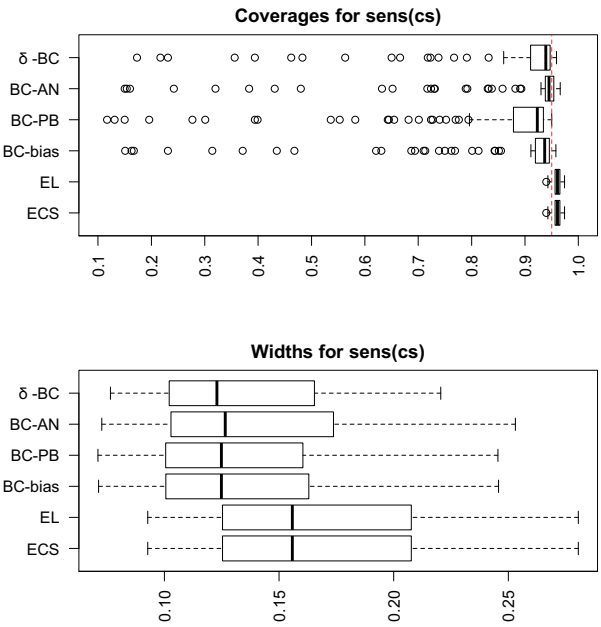

Figure 4. *Boxplot of the coverages and average widths of the confidence intervals for sens(c_s) for all scenarios combined.*

4. Application to SARS-CoV-2 antibody data

We use part of the dataset measuring the generation of antibodies elicited two weeks post-injection after the second dose of the BNT162b2 mRNA vaccine in adult healthcare workers (Kontopoulou et al. 2021). All subjects were closely followed on a quasi-daily basis in order to flawlessly detect the presence of COVID-19. The estimation of a cut-off point that could be used in medical practice would aid diagnosis of prior COVID-19 infection, which in turn can guide decision making regarding correct classification of patients with symptoms pertinent to post COVID-19 syndrome (long covid). The data consist of 289 subjects without prior COVID-19 and 50 subjects with confirmed prior COVID-19. Antibody data do not conform to normality assumptions in general (Anderson and Darling, 1952) normality test p-value $< 2.2e-16$ for controls and 0.002188 for cases) and a log10 transformation is a typical remedy in order to pursue formal hypothesis testing (Horne-Dale, 1995). The Box-Cox transformation provides a straightforward approach for an optimal estimation of the power transformation to normality. Kernel density estimators illustrating measurements before and after the Box-Cox transformation ($\widehat{\lambda} = 0.235$) are given in Figure 7 (a) and (b), respectively. In addition, we have included the corresponding estimation of the densities under the binormal model (after the Box-Cox transformation) and its corresponding ROC curve, whose associated AUC value is 0.716, in Figure 7 (c) and (d), respectively.

Results regarding the cut-off point based on the symmetry point methods presented in Section 2 are given in Table 1, suggesting that antibody measurements above around 22300 suggest prior COVID-19 infection, when antibodies are measured two weeks after the second dose of the BNT162b2 mRNA vaccine, with sensitivity around 65%. The classification of patients, with symptoms pertinent to post COVID-19/long Covid syndrome is to be taken with caution in practice. Antibody levels two weeks post-injection after the second dose of the BNT162b2 mRNA vaccine is a significant marker but cannot be used as a standalone test in practice given its very limited utility nowadays and its moderate performance. Corresponding confidence regions (CRs) and their illustration are given in Table 2 and Figure 8. Confidence regions are somewhat tighter using the R^*_α approach, providing an apparent higher estimation accuracy.

Table 1. *Estimates and* 95% *CIs for the different methods.*

Method	$\widehat{c}_s$	95% CI $(\widehat{c}_s)$	$\widehat{sens}(c_s) = \widehat{spec}(c_s)$	95% CI $sens(\widehat{c}_s)$
δ	25234.90	(22814.74, 27655.06)	0.641	(0.578, 0.703)
δ -BC	22304.13	(20103.32, 24504.93)	0.656	(0.601, 0.711)
BC-AN	22220.90	(19888.72, 24553.08)	0.656	(0.597, 0.714)
BC-PB	22299.64	(20103.53, 24495.76)	0.661	(0.608, 0.713)
BC-Bias	22251.90	(20119.51, 24384.29)	0.651	(0.599, 0.704)
EL	22296.64	(19597.31, 24995.98)	0.655	(0.591, 0.719)
ECS	22296.64	(19597.31, 24995.98)	0.655	(0.591, 0.719)

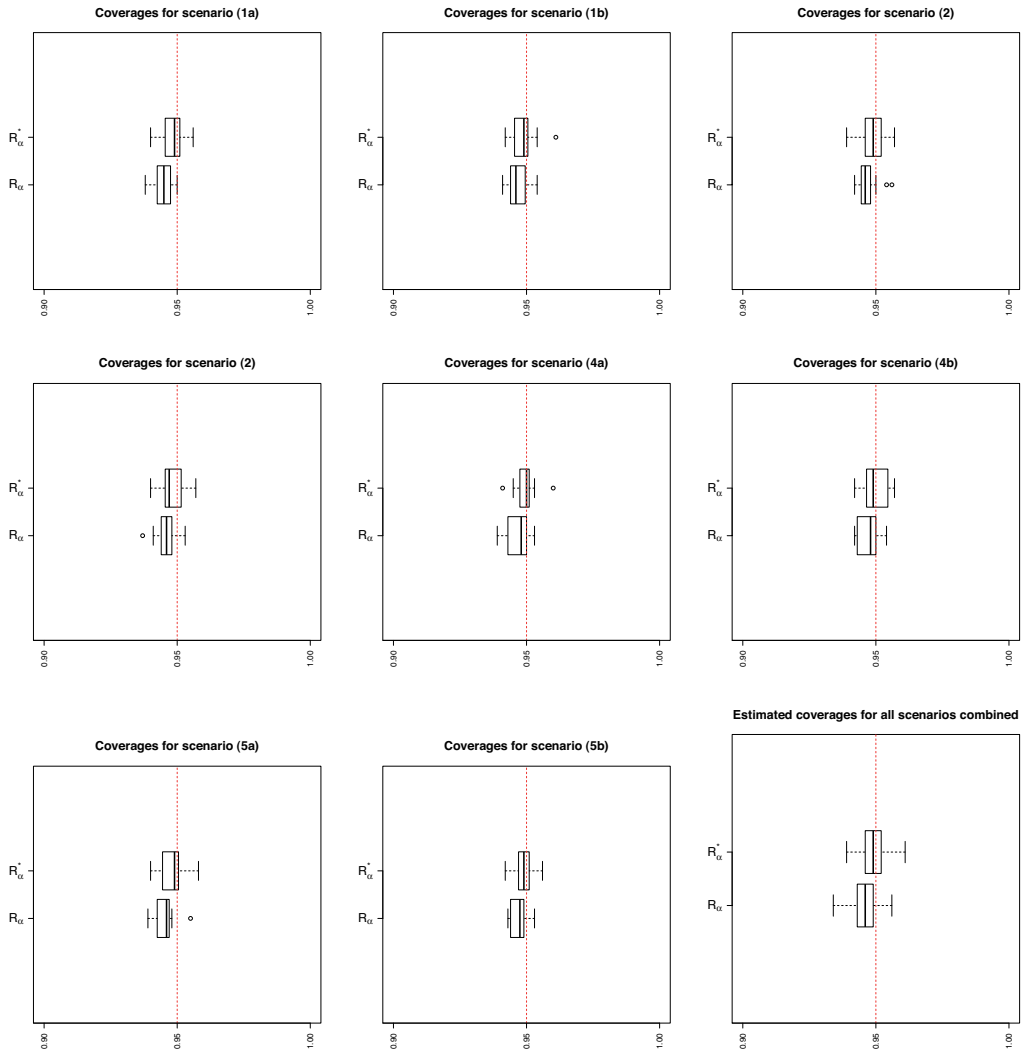

Figure 5. *Boxplots of the estimated coverages of the confidence regions at 95% confidence level for all scenarios considered in the simulation study, including a boxplot (the last one) with the results of all scenarios combined.*

Table 2. *Estimates of* $90, 95, 99\%$ *CRs for the different methods.*

Area	90% CR	95% CR	99% CR
R_α	463.62	621.10	946.14
R^*_α	459.47	603.13	939.89

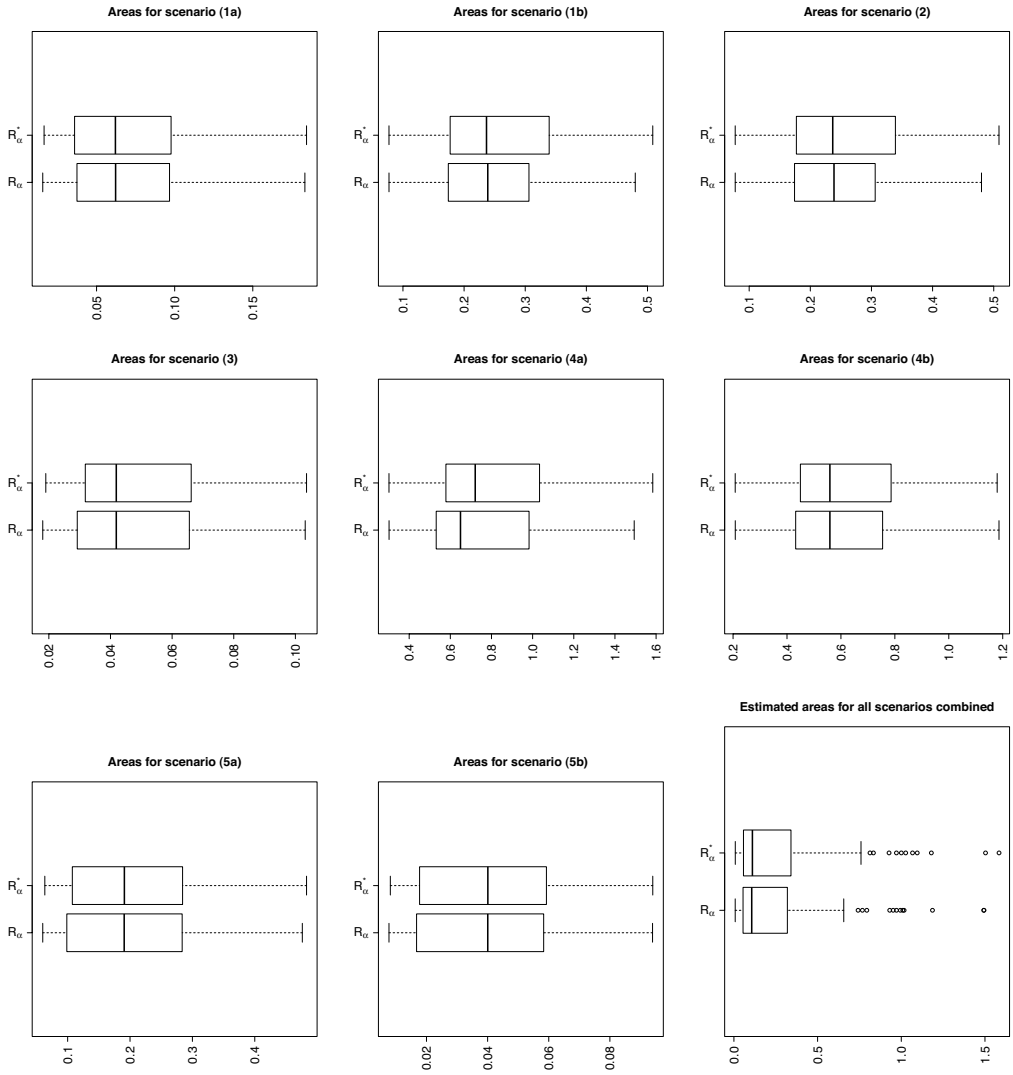

Figure 6. *Boxplot of the estimated areas of the confidence regions at 95% confidence level for all scenarios considered in the Simulation Study, including a boxplot (the last one) with the results of all scenarios combined.*

5. Discussion

When considering a continuous biomarker/score clinicians are in need of a cut-off point/ optimal threshold to classify subjects into one of the two diagnostic groups under consideration. One such method is based on the symmetry point. This approach has theoretical

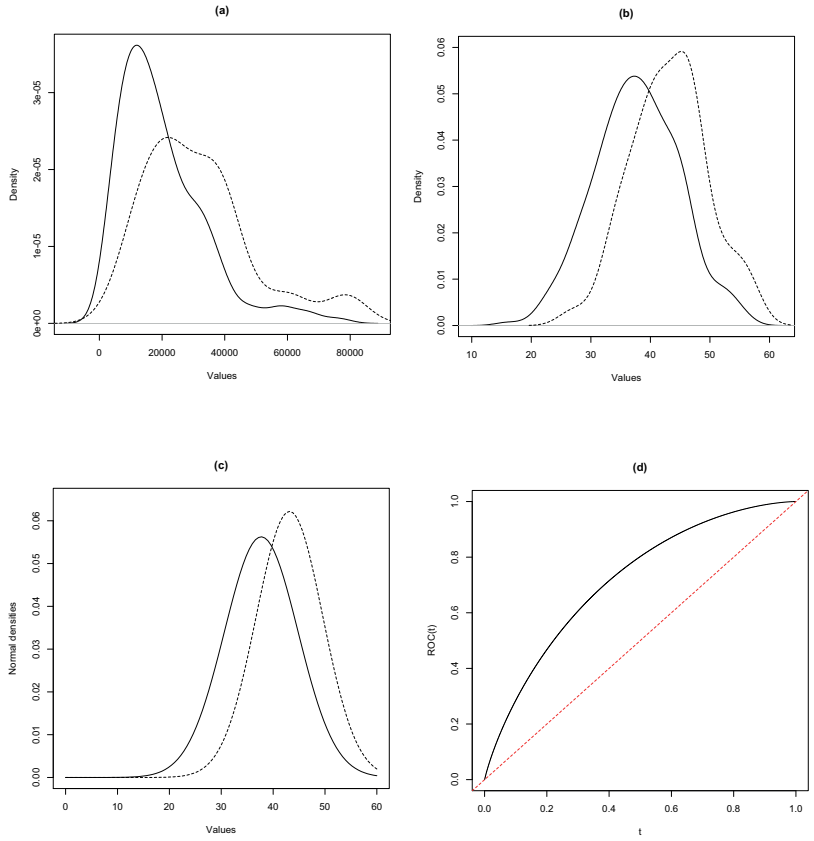

Figure 7. *Top row: Kernel density estimators before (a) and after (b) BC for the SARS-CoV-2 antibody data. Bottom row: Estimated densities (c) and corresponding ROC curve (d) under the binormal model, after BC.*

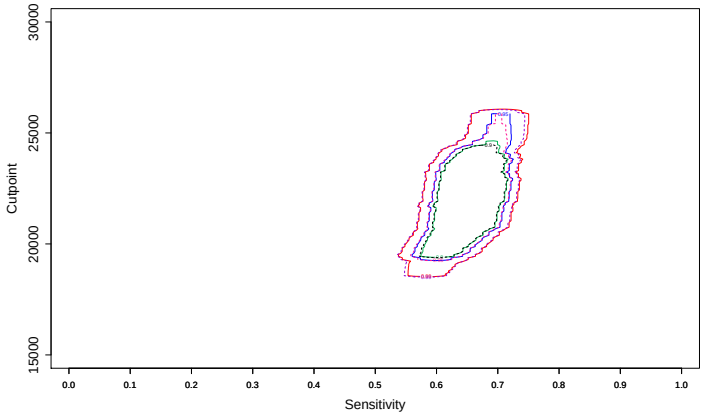

Figure 8. *CR contours for the SARS-CoV-2 antibody data. Solid lines correspond to R_α and dashed ones to R^*_α. Colours red and purple are used for the 99% CRs; blue and pink for the 95% CRs and green and black for the 90% CRs.*

support from a game theory minimax perspective and has important robustness features (Sanchez, 2017). The symmetry point has been used in applications but typically without the addition of confidence intervals. In this work, we have considered both marginal CIs for the symmetry point and its associated sensitivity and CRs for these jointly. For the marginal CIs we examined both parametric and nonparametric approaches. The parametric approaches are based on the binormal model and implemented using the Box-Cox transformation for cases when normality assumptions are not met. The BC-AN method worked quite well for the associated sensitivity for a wide range of distributions except for the mixtures. Two nonparametric methods were examined. The first due to López-Ratón et al. (2016) is based on empirical likelihood while the second is based on the chi-square test statistic and has been found to work well in other contexts. Both of these give very similar results and work quite well both for the symmetry point and its corresponding sensitivity although somewhat conservative for most scenarios. Due to the poor performance of the parametric CIs for c_s we only examined nonparametric confidence regions for the symmetry point and its associated sensitivity index. The first due to Adimari and Sinigaglia (2020) is based on empirical likelihood. We proposed an alternative again based on a chi-square test statistic. Our simulations indicate that although these two methods perform similarly our proposal generally provides results with observed coverage closer to the nominal level. The availability of CIs and CRs for the symmetry point approach should help practitioners using this method in their data analyses. Software for carrying out these procedures is available from the first author. We illustrated these procedures using part of the dataset from a published study on SARS-CoV-2 antibody levels post vaccination.

In addition to the symmetry point approach there are many other methods proposed in the literature for obtaining the optimal cut-off point value. Commonly seen methods include maximizing the Youden Index (Bantis, Nakas and Reiser, 2019) or its weighted version (Schisterman et al., 2005), point closest to the (0,1) corner (Perkins and Schisterman, 2006) and maximizing the product of sensitivity and specificity (Liu, 2012) among others. With many possible methods there is an inherent problem in choosing the appropriate method for the selection of an optimal cutoff point that can be used in everyday practice. Weights reflecting the relative importance of sensitivity and specificity can be introduced in a cost-benefit tradeoff approach. Researchers can be expected to potentially run a large number of cut-off point selection approaches and estimate cut-off points along with corresponding CIs prior to final decision making. It is difficult to say which approach is more clinically relevant. No automatic procedure for such a choice currently exists and it is not clear if this is possible at all. Future research on this issue would be beneficial for practical applications in clinical problems. An initial step in this direction would be a detailed comparison of cut-off point selection methods along with the assessment of the robustness of corresponding CIs.

Acknowledgments

This work was supported by grants PID2019-104681RB-I00. Data courtesy of Dr Konstantina Kontopoulou.

References

Anderson, T.W. and Darling, D.A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *Ann. Math. Statist.*, 23, 193-212.

Arnone, E., Cucchi M., Dal Gesso, S., Petitta, M. and Calmanti, S. (2020). Droughts Prediction: a Methodology Based on Climate Seasonal Forecasts. *Water Resources Management*, 34, 4313-4328.

Adimari, G. and Sinigaglia, A. (2020). Nonparametric confidence regions for the symmetry point-based optimal cutpoint and associated sensitivity of a continuous-scale diagnostic test. *Biometrical Journal*, 62, 1463-1475.

Bantis, L.E., Nakas, C.T. and Reiser, B. (2019). Construction of confidence intervals for the maximum of the Youden index and the corresponding cutoff point of a continuous biomarker. *Biometrical Journal*, 61(1), 138-156.

Box, G.E. and Cox, D.R. (1964). An analysis of transformations. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 26, 211-252.

Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine*, 21, 3093-3106.

Fluss, R., Faraggi, D. and Reiser, B. (2005). Estimation of the Youden Index and its Associated Cutoff Point, *Biometrical Journal*, 47, 458-472.

Franco-Pereira, A.M., Nakas, C.T. and Pardo, M.C. (2020). Biomarker assessment in ROC curve analysis using the length of the curve as an index of diagnostic accuracy: the binormal model framework. *AStA Advances in Statistical Analysis*, 104, 625-647.

Franco-Pereira, A.M., Nakas, C.T., B. Reiser and Pardo, M.C. (2021). Inference on the Overlap Coefficient: The binormal approach and alternatives. *Statistical Methods in Medical Research*, 30, 2672-2684.

Horne-Dale, A. (1995). The Statistical Analysis of Immunognicity Data in Vaccine trials. *Annals of the New York Academy of Sciences*, 754, 329-346.

Kontopoulou, K., Ainatzoglou, A., Nakas, C.T., Ifantidou, A., Goudi, G., Antoniadou, E., Adamopoulos, V., Papadopoulos, N., and Papazisis, G. (2021). Second dose of the BNT162b2 mRNA vaccine: Value of timely administration but questionable necessity among the seropositive. *Vaccine*, 39, 5078-5081.

Le, R., Ku, H. and Jun, D. (2021). Sequence-based clustering applied to long-term credit risk assessment. *Expert Systems With Applications*, 165, 113940.

Liu, X. (2012). Classification accuracy and cut point selection. *Statistics in Medicine*, 31(23), 2676-2686.

López-Ratón, M., Cadarso-Suárez, C., Molanes-López, E.M. and Letón, E. (2016). Confidence intervals for the symmetry point: An optimal cutpoint in continuous diagnostic tests. *Pharmaceutical Statistics*, 15, 178-192.

Molodianovitch, K., Faraggi, D. and Reiser, B. (2006). Comparing the areas under two correlated ROC curves: parametric and non-parametric approaches. *Biometrical Journal*, 48, 745-757.

Pardo, J.A. and Pardo, M.C. (2008). Minimum phi-divergence estimator and phi-divergence statistics in Generalized linear models with binary data. *Methodology and Computing in Applied Probability*, 10, 357-379.

Pardo, M.C., Lu, Y. and Franco-Pereira, A.M. (2022). Extensions of empirical likelihood and chi-squared-based tests for ordered alternatives. *Journal of Applied Statistics*, 49, 24-43.

Perkins, N.J. and Schisterman, E.F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, 163(7), 670-675.

Sanchez, IE. (2017) Optimal threshold estimation for binary classifiers using game theory [version 3; peer review: 3 approved] *F1000Research 2017*, 5(ISCB Comm J): 2762.

Sande, S.Z., Seng, L., Li, J. and D' Agostino, R. (2021). Statistical Learning in Medical Research with Decision Threshold and Accuracy Evaluation. *Journal of Data Science*, 19, 634-657.

Schisterman, E.F., Faraggi, D., Reiser, B. and Hu, J. (2008). Youden Index and the optimal threshold for markers with mass at zero. *Statistics in medicine*, 27, 297-315.

Schisterman, E.F., Perkins, N.J., Liu, A. and Bondell, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*, 73-81.

Sekgala, M.D., Opperman, M., Mpahleni, B. and Mchiza, Z.J.-R. (2022). Anthropometric indices and cut-off points for screening of metabolic syndrome among South African taxi drivers. *Front. Nutr.* 9:974749.

Information for authors and subscribers

Author Guidelines

SORT accepts for publication only original articles that have not been submitted simultaneously to any other journal in the areas of statistics, operations research, official statistics or biometrics. Furthermore, once a paper is accepted it must not be published elsewhere in the same or similar form.

SORT is an **Open Access** journal which **does not** charge publication **fees**.

Articles should be preferably of an applied nature and may include computational or educational elements. Publication will be exclusively in English. All articles will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board.

Submission of papers must be in electronic form only at our **RACO** (Revistes Catalanes en Accés Obert) submission site. Initial submission of the paper should be a single document in **PDF** format, including all **figures and tables** embedded in the main text body. **Supplementary material** may be submitted by the authors at the time of submission of a paper by uploading it with the main paper at our RACO submission site. **New authors**: please register. Upon successful registration you will be sent an e-mail with instructions to verify your registration.

The article should be prepared in **double-spaced** format, using a **12-point** typeface. SORT strongly recommends the use of its LaTeX template.

The **title page** must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75–100 words) followed by the keywords and MSC2010 Classification of the American Mathematical Society.

Before submitting an article, the author(s) would be well advised to ensure that the text uses **correct English**. Otherwise the article may be returned for language improvement before entering the review process. The article must also use inclusive language, that is, language that avoids the use of certain expressions or words that might be considered to exclude particular groups of people, especially gender-specific words, such as “man”, “mankind”, and masculine pronouns, the use of which might be considered to exclude women. The article should also inform about whether the original data of the research takes gender into account, in order to allow the identification of possible differences.

Bibliographic references within the text must follow one of these formats, depending on the way they are cited: author surname followed by the year of publication in parentheses [e.g., Mahalanobis (1936) or Rao (1982b)])]; or author surname and year in parentheses, without comma [e.g. (Mahalanobis 1936) or (Rao 1982b) or (Mahalanobis 1936, Rao 1982b)]. The complete reference citations should be listed alphabetically at the end of the article, with multiple publications by a single author listed chronologically. Examples of reference formats are as follows:

- Article: Casella, G. and Robert, C. (1998). Post-processing accept-reject samples: recycling and rescaling. Journal of Computational and Graphical Statistics, 7, 139–157.

- Book: Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). Bayesian Data Analysis, 2nd Ed. Chapman & Hall / CRC, New York.

- Chapter in book: Engelmann, B. (2006). Measures of a rating’s discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. (eds), The Basel Risk Parameters: Estimation, Validation, and Stress Testing. Springer, New York.

- Online article (put issue or page numbers and last accessed date): Marek, M. and Lesaffre, E. (2011). Hierarchical generalized linear models: The R package HGLMMM. Journal of Statistical Software, 39 (13). http://www.jstatsoft.org/v39/i13. Last accessed 28 March 2011.

Explanatory footnotes should be used only when absolutely necessary. They should be numbered sequentially and placed at the bottom of the corresponding page. **Tables and figures** should also be numbered sequentially.

Papers should not normally exceed about **25 pages** of the **PDF** format (**40 pages** of the format provided by the SORT **LaTeX** template) including all figures, tables and references. Authors should consider transferring content such as long tables and supporting methodological details to the online supplementary material on the journal's web site, particularly if the paper is long.

Once the article has positively passed the first review round, the executive editor assigned with the evaluation of the paper will send comments and suggestions to the authors to improve the paper. At this stage, the executive editor will ask the authors to submit a revised version of the paper using the SORT **LaTeX** template.

Once the article has been accepted, the journal editorial office will **contact the authors** with further instructions about this final version, asking for the source files.

Submission Preparation Checklist

As part of the submission process, authors are required to check off their submission's compliance with all of the following items, and submissions may be returned to authors that do not adhere to these guidelines.

1. The submitted manuscript follows the guidelines to authors published by SORT
2. Published articles are under a Creative Commons License BY-NC-ND
3. Font size is 12 point
4. Text is double-spaced
5. Title page includes title, name(s) of author(s), professional affiliation(s), complete address of corresponding author
6. Abstract is 75-100 words and contains no notation, no references and no abbreviations
7. Keywords and MSC2010 classification have been provided
8. Bibliographic references are according to SORT's prescribed format
9. English spelling and grammar have been checked
10. Manuscript is submitted in PDF format

Copyright notice and author opinions

SORT Statistics and Operations Research Transactions
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58 - 08003 Barcelona. SPAIN
Tel. +34-93.557.30.76 – Fax +34-93.557.30.01
sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form

SORT *(Statistics and Operations Research Transactions)*

Name _____

Organisation _____

Street Address _____

Zip/Postal code _____ City _____

State/Country _____ Tel. _____

Fax _____ NIF/VAT Registration Number _____

E-mail _____

Date _____

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)***) from now on

Annual subscription rates:

– Spain: €42 (4 % VAT included)
– Other countries: €46 (4 % VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4 % VAT included)
– Other countries: €17/issue (4 % VAT included)

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-557 30 01

Or by e-mail to:
sort@idescat.cat