Journal of Economic Analysis 2025 4(2) 91-155

Model-Based Poststratification of Measurements that Imperfectly Cover the Universe Studied: The Case of Postal Delivery Times.

Alain Bultez [a, *], Bert Seghers [b]

[a] *Louvain School of Management, Catholic University of Louvain, FUCaM, Mons, Belgium*

[b] *International Post Corporation, Brussels, Belgium*

ABSTRACT

The goal of this paper is to work out a *poststratification* method for estimating *in-time*[1] *indicators* for *international end-to-end delivery processes* when

- collected data cannot cover all the strata making up the logistics universe to be surveyed,
- the true weights of the strata - needed to correct biases in representativeness caused by disproportionate sampling and incomplete coverage - are unknown but can be inferred from marginal subtotals related to stratification criteria considered separately, rather than jointly, and conditional on each end of the delivery journey: outbound- versus inbound-specific[2]. Within this perspective, *poststratification* is used here to mean a statistical correction of measurements derived from *incomplete stratified samples*, an *ex-post calibration* aimed at yielding more accurate estimates based on an analysis of the data. Thus, we tackle instances where *ex-ante* assignment to strata is not a problem, but when surveying all strata is out of the question.

[1] The ***"On-Time In-Full"*** (***OTIF***) supply chain metric is pertinently named to signal the extent to which customers receive exactly what they ordered on the agreed date. However, in practice, *"on time"* and *"in time"* are considered synonymous, whereas semantically they are not. In the following text, *"in time"* is used to connote: *"not late"* and *"punctuality"* for the fact of arriving on or before the scheduled time.

[2] In postal jargon, for international exchanges, *outbound* (alt., *inbound)* logistics refers to mail collection (alt. delivery) operations from the sender (alt., to the addressee) in the country of origin (alt., destination).

* Corresponding author: Alain Bultez
E-mail address: alain.bultez@uclouvain.be

ISSN 2811-0943
doi: 10.58567/jea04020006

Received 3 November 2024; Accepted 23 December 2024; Available online 13 February 2025; Version of Record 15 June 2025

For that purpose, an econometric model is designed

- to link the *discrete transport lead times, counted in days*, of tested items to the specifics of their material characteristics (e.g. size/weight), as well as those of the routes they take through the distribution network (e.g. origin and destination zones),
- and provide performance predictions for each of the strata, covered as well as non-covered.

Benchmarking the *multinomial cumulative logit* regression against the *negative binomial* one reveals that delivery time had better be treated as an *ordinal categorical system's response*, rather than as a ratio-scaled *count*.

The model-based fitted and extrapolated estimates are then used as inputs to the *ex-post weighting* stage, which produces robust point- and interval-estimates of aggregate *key performance indicators (KPIs)* through *bootstrapping*. Simple *linear programs* provide two extreme weighting sets, one per country-to-country path: the first minimizes the *KPIs'* values, while the second maximizes them.

Probabilities of delivery within deadlines summarize distributions of delivery times better than their means and standard deviations, because logistical efforts to cut transit by one day must be enhanced more and more as it gets shortened. Three types of graphs are proposed to help visualize this exponential increase in the service quality required. The applicability of the methodology developed is demonstrated on the 2023 database of the *International Post Corporation*. In this case, the imprecision of the *KPI* estimates depends much more on the uncertainty caused by disturbances occurring during the *first-* and *last-miles*[3], than on the imperfection of the information about the real weights of the strata.

KEYWORDS

End-to-End Logistics; Lead Time; Probability of Delivery Within Deadline; Incomplete Coverage; Poststratification; Ordinal Categorical Response; Count Variable; Multinomial Cumulative Logit Regression; Negative Binomial Regression; Linear Programming; Bootstrapping

[3] In the broadest sense the transportation industry attaches to these words: initial and final *legs/stretches* of the route.

1. Introduction

Surveying end-to-end[4] multi-firm cross-border distribution channels by testing the routing of items streamed through them is a complex task. Indeed, such universes are too vast and multi-sided to be fully covered and sampled proportionately. Encompassing all classes of objects carried, and every path they might pass through, is just unfeasible. Numerous strata remain unobserved and many more are barely sampled at all. To correct for biases caused by the resulting sample unrepresentativeness, missing measurements must be conjectured before applying any statistical redressing. On the data collected on covered strata, an econometric model - using the stratification criteria as predictors - can first be calibrated to infer measurements for the non-covered ones. Next, to obtain consistent aggregate estimates, the model-predicted strata-level measurements - either extrapolated (for non-covered strata) or fitted (for covered strata) - need to be weight-averaged, according to the relative strata sizes in the universe.

The present paper is probably the first to be published on the development such a comprehensive methodology and proving its applicability to the quality control of international transports, more specifically: postal services. The only related earlier work one can mention is the article of Caulkins et al. (1993) who formalized and compared variants of a scoring two-factor[5] multiplicative equation to rate simultaneously the *US* domestic airlines' *promptness* and the accessibility of the airports to which they fly, exploiting secondary data issued by the *Department of Transportation*: percentages of on-schedule landings, per company and per airport. Although they cared much about the fairness of the marks their model awards, they did not assess their statistical accuracy.

Section **2** complements the present introduction by setting more precisely the stage: the cross-border delivery of the priority letter mail across Europe, tracked by the ***International Post Corporation*** (***IPC***), pursuant to the Directive 97/67/EC of the European Parliament and of The Council of 15 December 1997 on *"common rules for the development of the internal market of Community postal services and the improvement of quality of service"*[6].

Section **3** defines the problematic to be faced: the multidimensionality of the postal world to be supervised, the *in-time* indicators to be coursed, and the uncertainty about the real strata weights in the postal universe. In addition, it proposes a linear programming solution to deal with the resulting indeterminacy of key performance statistics.

Section **4** goes over econometric specifications relevant for quantifying relationships between delivery times and their potentially determining factors. Thus, it reviews the literature on *count* and *logistic regression* models and shows how to adapt them to predict shipping times.

Section **5** compares outputs from implementing both classes of models to the 2023 ***IPC*** data base: the *negative binomial* and the *multinomial cumulative logit*. It argues that delivery time should be treated as a *categorical ordinal response*, rather than as a *ratio* (or *interval*) measure.

Section **6** demonstrates how the *bootstrapping* technique can be used to derive robust *confidence intervals* for probabilities of delivery within t working days of the *collection date:* D. It completes the case by commenting on the most critical in-time delivery rates: i.e., the percentages delivered to addressees no later than $D + 5$, the upper time limit set by the *EU "for postal items of the fastest standard category"*[7].

Section **7** concludes by discussing the limits of our approach, suggesting future lines of research, and highlighting the strengths of our contribution.

---

[4] *"End-to-end routing is measured from the access point to the network to the point of delivery to the addressee"* (endnote: *, L 15/25, in https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31997L0067).

[5] In their own terms: *"the conditional probability that a flight on airline i will arrive on-time given that there are no local delays at the airport of arrival"* and *"the unconditional probability that a flight arriving at airport j will encounter no local delays"* (op. cit., p. 713).

[6] Cf. source referred to in footnote 4.

[7] Cf. the source referred to in footnote 4: *Annex*, L 15/25.

All computer codes programmed in ***SAS (Statistical Analysis Software)***, to carry out the research work reported hereafter, are listed and commented at length in the **Appendix**.

2. Heart of the matter: effectiveness of postal logistics from the end-customer's viewpoint

Timely, careful and cost-effective cross-border transportation of products demands intense, transparent and loyal collaboration from all parties involved in their routing. Therefore, the logistics of various Posts, who may fiercely compete against one another on domestic markets, must be aligned. Their mutual aid is called for to ensure that

- mail collected from a sender in the origin country by a first player is not only efficiently carried up to its outbound handover-point,
- but ultimately gets delivered by a second one from the latter's inbound handover-point to the addressee in the destination country.

By sharing a common systemic supply-chain vision, rivals become partners at least for their international service-lines. In this spirit, aware of the technological challenges ahead and convinced that a cooperative game is preferable to a zero-sum one, in 1989, the industry majors allied by joint-venturing into the *International Post Corporation* (***IPC***), which promotes the four coordination modes designed to streamline supply chains: *"logistics synchronisation, information sharing, incentive alignment, and collective learning"* (Simatupang et al., 2002, pp. 291-301).

2.1. International Post Corporation

IPC-was established as a cooperative association positioning itself as *"the leading service provider of the global postal industry that provides leadership by driving service quality, interoperability and business-critical intelligence to support Posts in defending existing business and expanding into new growth areas"* (https://www.ipc.be/).

It currently has 26 members, from Posts operating in the Asia-Pacific region, Europe and North America. ***IPC*** services are used globally by over 180 posts worldwide. It provides platforms, programs and tools for their CEOs and senior managers to share best practice and discuss strategy. It also conducts market research on the industry and manages a system of incentive payments between operators for their delivery activities. Acting as both a carrot and a stick, ***IPC***-driven remuneration agreements ensure that the focus is on quality-of-service (***QoS***) and foster data exchanges for tracked products. ***IPC*** also supports specific e-commerce endeavours designed to facilitate product returns and enhance the sustainability of postal logistics. Above all it actively promotes quality control initiatives. Since its inception, it has applied high standards to upgrade logistics procedures and performance. To this end, it has developed technological solutions to monitor and improve the ***QoS*** for international letters, packets and parcels. By scanning barcoded products at crucial stages of their end-to-end postal journey, ***IPC*** records events and pinpoints bottlenecks that may encumber their handling. Since such in-process information is unavailable for non-barcoded items – letters, and small e-commerce packets, called: *"untracked mail"* – ***IPC*** built the ***UNEX™ QoS***-measurement.

2.2. UNEX™

The name ***UNEX™*** results from the concatenation of the first syllables of two terms: first, ***UN****IPOST* meant to evoke joined forces amongst postal organizations and ***EXternal*** to signify that its methodology is independently regulated and audited by third parties, external to the Posts, to minimize risks of conflict of interests. ***UNEX™*** is ***IPC*** core system quantifying the ***QoS*** provided to transport international single-piece priority letter mail. Using a mystery shopper approach, it mobilizes volunteer *"panellists"*, coached to act like ordinary customers sending or

receiving test items copycatting real mail. Monitoring closely the tour of the ***UNEX™*** test pieces through the postal pipeline yields transit-times' observations. To secure integrity and representativeness of such data all volunteer panellists are recruited, trained and supervised by an external market-research agency whose activities are hidden from ***IPC*** and their member Posts. All ***UNEX™*** items are prepared centrally by this contractor, taking into account country-specific habits and national postal requirements in the countries of origin and destination, to ensure that these test items are indistinguishable from actual postal items processed between these countries. These test-letters are then dispatched to the sender-panellists, with precise instructions about when and where they are to be inducted into the postal flows.

Each ***UNEX™*** envelope contains a ***r***adio ***f***requency ***id***entifier, also known as an ***RFID*** tag or transponder: a small tracking device in sticker form that records its passage at critical points along its journey through the end-to-end postal chain (from collection to delivery). This enables Posts to break down the logistics chain into stretches and so to bring much more detailed ***QoS*** signals to better target operational steps within the postal network and design specific action plans and corrective measures.

Detailed feedback from panellists is collected via an online platform. The ***UNEX™*** software package checks the consistency of every registration against guidelines, as well as in relation to the events tracked by the transponders (e.g., by ascertaining whether deliveries have exceptionally taken place over a weekend). The exact time the cross-validated test-item spent within the postal circuit is determined and depending on whether or not it has arrived within the prescribed deadline, it will be counted as ***'in time'*** or ***'late'***.

To fulfil specific requirements of various Posts and internal stakeholders (postal communities), ***IPC*** runs ***UNEX™*** modules tailored to three objectives: regulation, operations and incentive-based remunerations. The operations modules produce detailed information on the processing of the test letters so that Posts can identify bottlenecks or weak spots, undertake root-cause analyses and follow-up with action plans. The ***UNEX™ TD*** module supplies independent and reliable ***QoS*** reports which assist Posts in setting the amounts of terminal dues to be exchanged, according to performances in compliance with agreed service standards. However, this article exclusively deals with the regulatory facet explained under 2.3.

*2.3. **UNEX™ CEN** module*

With the opening up of the postal European market on 15 December 1997 - marking the end of the postal operators' monopoly for cross-border single piece priority mail in many countries -, Posts are duty-bound to report on how they are fulfilling the universal service obligation imposed by the ***EU*** Postal Directive[8]. The injunctions specifying how to assess this universal service within Europe were detailed in the ***CEN*** standard: ***EN 13850:2020***[9]. The ***CEN*** standard serves as a kind of biblical compass, defining for European countries, regulatory postal ***QoS*** measurements on letter mail, the methodology to be followed, constraints to be respected, and degrees of freedom left, to

- set up the ***stratification design***, a necessary condition for a robust measurement,
- discriminate in-time (coded:**1**) test-items from delayed ones (coded: **0**),
- aggregate the item-specific in-time binary indicators into representative and reliable proportions of successes, appraise the precision of these aggregate estimates of in-time delivery probabilities.

Figure 1 maps the Posts tested via the ***UNEX™-CEN*** module in 2023.

[8] Directive: https://eur-lex.europa.eu/eli/dir/1997/67/oj, and successive amendments: https://eur-lex.europa.eu/eli/dir/2002/39/oj, https://eur-lex.europa.eu/eli/dir/2008/6/oj.

[9] ***CEN*** stands for ***Centre Européen de Normalisation*** (https://www.cencenelec.eu/about-cen). Organized in Technical Committees (TC), ***CEN*** contributes to the development of European standards and releases technical documents relating to various types of products, materials, services and processes. The postal services are covered by TC 331 focusing on the standardization of various aspects of the ***QoS*** measurement, in order to increase the interoperability of postal networks and thereby improve the ***QoS***.

Figure 1. Postal territories and operators.

In application of the standard, ***IPC*** annually publishes postal ***QoS***-levels reached in Europe[10]. Also, the ***UNEX™-CEN*** module gets regularly audited to certify its conformance to the ***CEN***-norm.

For 2023, ***IPC*** monitored 132,304 items, out of which **105, 889** records were screened and found valid. They were sent and/or received by 2,522 volunteer panellists, covering **710** country-to-country flows within Europe. Field studies are delimited by the *country-to-Europe (Co2Eu)* and *Europe-to-country (Eu2Cd)* streams to ensure that each national regulatory authority can monitor their territory from aggregate *inbound* and *outbound* outlooks. Therefore, all information needed to build the statistical design, and the sampling plan, is collected at the level of the whole country and is not broken down across individual *country-to-country (Co2Cd)* flows. Although *RFID*-tags allow the end-to-end process to be split into its detailed paths, ***UNEX™ CEN*** is solely concerned by the end-to-end in-time proportions from the customer perspective.

[10] https://www.ipc.be/services/operational-performance-services/unex/results./

3. Challenge: multidimensionality

Table 1 defines the typology of test-items. It lists their features.

Table 1. Coding of the discriminant mail characteristics.

Fixed Nominal Factors Each identified by a superscript: $f \in \mathcal{F}$. Set of modes: $\boldsymbol{M^f}$. Cardinality: $\left\|\boldsymbol{M^f}\right\|$.	Classification Dummies Binary variables: $x^{f}_{i,m(f)}$.
Outbound logistics in origin country: $f = Co$ $\boldsymbol{M^{Co}} = \{AT, \dots, SK\}$, set of ISO3166 alpha-2 codes, with: $\left\|\boldsymbol{M^{Co}}\right\| = \mathbf{31}$.	$x^{Co}_{i,m(Co)} = \begin{cases} 1, \text{if } i \text{ sent from country } m(Co), \\ 0, \text{otherwise,} \end{cases}$ for $m(Co) \in \boldsymbol{M^{Co}}$.
Urbanization of origin-area: $f = Uo$ $\boldsymbol{M^{Uo}} = \{Ca$: Capital, Kc: Key cities, Rt: Rural towns$\}$, with: $\left\|\boldsymbol{M^{Uo}}\right\| = \mathbf{3}$.	$x^{Uo}_{i,m(Uo)} = \begin{cases} 1, \text{if } i \text{ sent from a } m(Uo) \text{ area,} \\ 0, \text{otherwise,} \end{cases}$ for $m(Uo) \in \boldsymbol{M^{Uo}}$.
Size/weight of the envelope: $f = Sw$ $\boldsymbol{M^{Sw}} = \{B1\_50g; C5\_50g, C6\_20g\}$, with: $\left\|\boldsymbol{M^{Sw}}\right\| = \mathbf{3}$.	$x^{Sw}_{i,m(Sw)} = \begin{cases} 1, \text{if } i \text{ of size/weight } m(Sw), \\ 0, \text{otherwise,} \end{cases}$ for $m(Sw) \in \boldsymbol{M^{Sw}}$.
Address labelling: $f = Ad$ $\boldsymbol{M^{Ad}} = \{Mt$: Machine typed; Hw: Handwritten$\}$, with: $\left\|\boldsymbol{M^{Ad}}\right\| = \mathbf{2}$.	$x^{Ad}_{i,m(Ad)} = \begin{cases} 1, \text{if } i\text{'s adress labelling is } m(Ad), \\ 0, \text{otherwise,} \end{cases}$ for $m(Ad) \in \boldsymbol{M^{Ad}}$.
Franking: $f = Fk$ $\boldsymbol{M^{Fk}} = \{St: Stamped; Mt: Metered; Pp: prepaid\}$, with: $\left\|\boldsymbol{M^{Fk}}\right\| = \mathbf{3}$.	$x^{Fk}_{i,m(Fk)} = \begin{cases} 1, \text{if } i\text{'s franking is } m(Fk), \\ 0, \text{otherwise,} \end{cases}$ for $m(Fk) \in \boldsymbol{M^{Fk}}$.
Sending weekday: $f = Wd$ $\boldsymbol{M^{Wd}} = \{Mo: Monday, .., Sa: Saturday\}$, with: $\left\|\boldsymbol{M^{Wd}}\right\| = \mathbf{6}$.	$x^{Wd}_{i,m(Wd)} = \begin{cases} 1, \text{if } i \text{ sent on weekday } m(Wd), \\ 0, \text{otherwise,} \end{cases}$ for $m(Wd) \in \boldsymbol{M^{Wd}}$.
Induction place: $f = Pl$ $\boldsymbol{M^{Pl}} = \{Sb: Streetbox, Po: Post\ office, Pu: Pickup\}$, with: $\left\|\boldsymbol{M^{Pl}}\right\| = \mathbf{3}$.	$x^{Pl}_{i,m(Pl)} = \begin{cases} 1, \text{if } i \text{ inducted in } m(Pl), \\ 0, \text{otherwise,} \end{cases}$ for $m(Pl) \in \boldsymbol{M^{Pl}}$.
Inbound logistics in destination country: $f = Cd$ $\boldsymbol{M^{Cd}} = \{AT, \dots, SK\}$, , set of ISO3166 alpha-2 codes, with: $\left\|\boldsymbol{M^{Cd}}\right\| = \mathbf{32}$.	$x^{Cd}_{i,m(Cd)} = \begin{cases} 1, \text{if } i \text{ sent to country } m(Cd), \\ 0, \text{otherwise,} \end{cases}$ for $m(Cd) \in \boldsymbol{M^{Cd}}$.
Urbanization of destination-area: $f = Ud$ $\boldsymbol{M^{Ud}} = \{Ca$: Capital, Kc: Key cities, Ra: Rural towns$\}$, with: $\left\|\boldsymbol{M^{Ud}}\right\| = \mathbf{3}$.	$x^{Ud}_{i,m(Ud)} = \begin{cases} 1, \text{if } i \text{ sent to a } m(Ud) \text{ area,} \\ 0, \text{otherwise,} \end{cases}$ for $m(Ud) \in \boldsymbol{M^{Ud}}$.
Maximum number of strata: $\boldsymbol{TNS} = \prod_{f \in \mathcal{F}} \left\| \boldsymbol{M^f} \right\| = \mathbf{2,892,672}$	Number of independent dummies: $\sum_{f \in \mathcal{F}} \left(\left\| \boldsymbol{M^f} \right\| - 1 \right) = 77$

Indeed, the ***UNEX™ CEN***-module evaluates the ***QoS*** offered by Posts for the end-to-end transfer of priority letters

- of sizes/weights: $B1\_50g, C5\_50g, C6\_20g$,
- *franked* via *stamps*, *metering machine marks* or *pre-paid envelopes*,
- *sent* on either *Monday, Tuesday, …., or Saturday*,
- from collection points located in either the *capital*, or a *key-city*, or *rural town*,
- in the *origin country*: *street boxes, postal counters* or *post offices* and *in-situ pickup,*
- to the recipient living in either the *capital*, or a *key-city*, or a *rural town* in the *destination country*.

3.1. Huge and compound universe

The complexity of the postal logistic world to be monitored can best be grasped by delineating its *stratification*. The first column of Table 1 lists the *fixed factors* $(f \in \boldsymbol{\mathcal{F}})$ regarded as *predictors* of priority cross-border mail delivery punctuality, hence, to be used as partitioning dimensions, according to their *modes* $\big(m(f)\big)$.

3.1.1. Stratification criteria

These *fixed factors* consist of all the relevant shipping-related features that characterize each item to be handled:

- its material nature: size/weight (Sw), address labelling (Wd), franking (Fk),
- and its sender-to-receiver journey: outbound (Co) and inbound (Cd) countries, urbanization of areas where collection and delivery endpoints are located (Uo, Ud), sending weekday (Wd).

All together they form the set $\boldsymbol{\mathcal{F}} = \{Co, Uo, Sw, Ad, Fk, Wd, Pl, Cd, Ud\}$ including the nine factors deemed *discriminant* by the experts commissioned by the *European Committee For Standardization* (*CEN,* 2020, section 6.4, pp. 21-22), *"under the mandate given to CEN by the European Commission and the European Free Trade Association, and supports essential requirements of EU Directive(s)"* to police postal cross-border *Quality-of-Service* (***QoS***) measurement systems (Ibidem, p. 3).

Combining the *modes* of these nine traits yields the *stratification tree*, which in ***UNEX^{TM}-CEN*** counts largely more than two million branches, since theoretically the total number of *strata* (TNS) results from the product of the cardinal numbers $(|\ \ |)$ of the sets of modes $\big(|\boldsymbol{M^f}|, \forall f \in \boldsymbol{\mathcal{F}}\big)$:

$$TNS = \prod_{f \in \boldsymbol{\mathcal{F}}} |\boldsymbol{M^f}| = 2{,}892{,}672.$$

However, the actual number of strata (ANS) lies much lower, because many of them need to be assigned a zero weight because they do not exist. Indeed, the mode weights of the stratification criteria, other than outbound and inbound countries, vary according to either the origin or the destination of the mail.

3.1.2. Strata weighting

The information needed to design the testing is collected and processed at the country-to-Europe $(Co2Eu)$ and Europe-to-country $(Eu2Cd)$ levels (i.e. the fields of study), but it is operationalized in the measurement of country-to-country $(Co2Cd)$ links. In fact, ***IPC*** infers weights of modes (Ω_m) separately, one factor at a time, for[11]

- factors related to origin-countries - i.e., $f|Co \in \{Uo, Sw, Ad, Fk, Wd, Pl\}$ - from the distributions of these modes in the real mail sent from every country of origin, i.e., per Co, hence denoted by $\Omega_{m(f|Co)}$;
- factors related to destination-countries - i.e., $f|Cd \in \{Ud\}$ - from their distributions in the real mail sent to every country of destination, i.e., per Cd, hence denoted by $\Omega_{m(f|Cd)}$.

Therefore, the subset of the strata formed by the items carried from origins in Co to destinations in Cd - $S(Co2Cd)$ - is defined by the following conjunctions $(\cap)$ of shipments characteristics:

$$S(Co2Cd) \equiv \{[(Uo \cap Sw \cap Ad \cap Fk \cap Wd \cap Pl)|\boldsymbol{Co}] \cap [Ud|\boldsymbol{Cd}]\}.$$

The upper part of Table 2 details (up to and including *Output #1*) how ***IPC*** has so far implemented the ***CEN*** guidelines for the computation of the *conditional stream-specific strata weights* $\big(\omega_{s|Co2Cd}\big)$: thus, specific for every country-to-country $Co2Cd$-flow, i.e., for $s \in S(Co2Cd)$, and for all Co & $Cd \neq Co$ since domestic mail is out of the scope of ***UNEX^{TM}***. Formula (T2.1) determines what ***CEN*** calls the *standard weighting basis* (*SWB*) and recommends "*to define a proportional sample design"* (op. cit., §H.4.1.3, p. 88).

[11] The notation is deliberately general, to allow extensions to encompass additional attributes and factors.

Table 2. Formalization of the standard weighting basis.

Input #1. Conditional, outbound-specific, weight of mode $m(f|Co)$ of criterion $f|Co$:
$\Omega_{m(f|Co)}$, for $f|Co \in \{Uo, Sw, Ad, Fk, Wd, Pl\}$

Input #2. Conditional, inbound-specific, weight of mode $m(f|Cd)|$ of criterion $f|Cd$:
$\Omega_{m(f|Cd)|}$, for $f|Cd = Ud$.

Mode indicator. Binary variable indicating by which mode stratum s is characterized for every criterion:

$$I^{s}_{m(f|C)} = \begin{cases} 1, \text{if } s \text{ is characterized by mode } m(f|C) \text{ of criterion } f|C, \\ 0, \text{otherwise,} \end{cases} \quad \text{for } C \in \{Co, Cd\}.$$

Output #1. Conditional, $Co2Cd$ stream-specific, weight of stratum s:

$$\omega_{s|Co2Cd} = \prod_{C \in \{Co,Cd\}} \left(\prod_{f|C} \left\{ \sum_{m(f|C)} \left[I^{s|Co2Cd}_{m(f|C)} \times \Omega_{m(f|C)} \right] \right\} \right), \text{ for } s \in S(Co2Cd), \qquad \text{(T2.1)}$$

⇑ ⇑ ⇑

⇑ ⇑ {Selection of the conditional weight corresponding to the mode characterizing $s|Co2Cd$, w. r. t. $f|C$}

⇑ ⇑

Product over all criteria (factors) conditioned by C

⇑

Product of the two products of weights:
(a) product relevant to the factors related to the outbound end: $C = Co$,
(b) product relevant to the factors related to the inbound end: $C = Cd$.

Sum constraint, holding for every *Co2Cd* path : $\sum_{s \in S(Co2Cd)} \omega_{s|Co2Cd} = 1$.

Output #2. Marginal distribution of weights for the overall ***Eu2Eu***-*KPI*.
The marginal stratum weight $(\omega^{**}_{s|Co2Cd})$ results from
on the one hand, the within-path stratum conditional weight $(\omega_{s|Co2Cd}$: *Output* #1$)$,
and on the other, the share of the total mail volume flowing along that path, demoted: $\mathbb{v}_{Co2Cd}$. Thus,

$$\omega^{**}_{s|Co2Cd} = \omega_{s|Co2Cd} \times \mathbb{v}_{Co2Cd}, \text{when } s \in S(Co2Cd), \forall (Co, Cd \neq Co). \qquad \text{(T2.2)}$$

Overall sum constraint: $\sum_{o2d} \sum_{s \in S(o2d)} \omega^{**}_{s|o2d} = \sum_{o2d} \mathbb{v}_{o2d} \cdot \left[\sum_{s \in S(o2d)} \omega_{s|o2d} \right] = 1$, where: $\sum_{o2d} \equiv \sum_{o} \sum_{d \neq o}$.

However, according to the *SWB*, $\omega_{s|Co2Cd}$ results from the product of the weights of the modes that specify stratum s, which is only valid if the distributions of modes are independent from one another across characteristics. ***CEN*** underlines that such an assumption may not hold, noting a counterexample: stamped envelops $(m(Fk) = St)$ are less frequently inducted at the post office $(m(Pl) = Po)$ than metered ones $(m(Fk) = Mt)$. Regrettably, given the lack of detailed information on the split of mail volumes per mix of characteristics, ***IPC*** has stuck to the *SWB*; however below, under *3.4*, we revisit this limitation and explain how it can be relaxed.

The second part of Table 2 (under Output #2: formula (T2.2)) exposes how the *SWB* conditional, stream-specific, strata weights should be adjusted - weighting them in turn by the $Co2Cd$ *mail volume-shares* $(\mathbb{v}_{Co2Cd})$ - to obtain the *marginal* weights: $\omega^{**}_{s|Co2Cd}$, for $s \in S(Co2Cd)$, needed to evaluate the *overall end-to-end QoS* provided jointly by all postal operators involved in mail delivery within Europe.

3.2. Ex-post weighting, necessity for sound diagnoses

To warrant unbiased estimates from straightforward statistical processing, samples must be allocated proportionally to strata weights. However simple this type of plan is in principle, it is hard, if not impossible, to execute because of all sorts of constraints (e.g., budget, timing, …) and field contingencies that disturb implementation (e.g., panelists failing to comply with instructions). And of course, the more sophisticated the experimentation blueprint, the more unfeasible it is. Moreover, disproportionate sampling may yield more accurate estimates: oversampling (under-sampling) strata for which performances are more (less) uncertain is indeed desirable, provided an *ex post corrective weighting* restores proportionality to strata relative importance in the universe. For sure, the international postal universe is so intricate that a full scale proportional experimental design would come to a dead-end. Therefore, strata-level records need to be *weighted.* More precisely, sampling imbalances must be redressed, when compiling *strata-level estimates of probabilities of delivery within a fixed deadline of* $\mathbb{t}$ *periods* denoted by $\widehat{\Pi}_{s,\mathbb{t}}$. Indeed, correctly weighted arithmetic means of these $\widehat{\Pi}_{s,\mathbb{t}}$, - made explicit in the next two paragraphs - yield valid aggregate indicators of logistic effectiveness at different echelons.

3.2.1. ***IPC*** scorecard

At the lowest echelon, ***IPC*** can assess the postal logistics for every country-to-country route ($Co2Cd$), by an estimate of the probability of delivery within a $\mathbb{t}$-day deadline of any item sent from within Co to an addressee in Cd, denoted by $\widehat{\boldsymbol{\Pi}}(\boldsymbol{Co2Cd},\mathbb{t})$**.** The latter results from the weighted average, over the subset: $s \in S(Co2Cd)$, of the $\widehat{\Pi}_{s,\mathbb{t}}$, using the conditional weights defined by formula (T2.1):

$$\widehat{\boldsymbol{\Pi}}(\boldsymbol{Co2Cd},\mathbb{t}) = \sum_{s\in S(Co2Cd)} \left[\omega_{s|Co2Cd} \times \widehat{\Pi}_{s,\mathbb{t}}\right], \forall (Co, Cd \neq Co) \tag{1}$$

At the top of its dashboard, ***IPC*** places the supra-aggregate *KPI* summing up the overall *QoS* of all the country-to-country flows. This ultimate *KPI* can be obtained by extension of (1) to all $o2d$-routes, substituting the $\omega^{**}_{s|o2d}$ - calculated by (T2.2) - for the $\omega_{s|o2d}$:

$$\widehat{\boldsymbol{\Pi}}(\boldsymbol{Eu2Eu},\mathbb{t}) = \sum_{o2d}\left(\sum_{s\in S(o2d)}\left[\omega^{**}_{s|o2d} \times \widehat{\Pi}_{s,\mathbb{t}}\right]\right) = \sum_{o2d}\sum_{s\in S(o2d)}\left[\left\{\omega_{s|o2d} \times \mathbb{v}_{o2d}\right\} \times \widehat{\Pi}_{s,\mathbb{t}}\right]. \tag{2}$$

3.2.2. Outbound versus inbound trackers

To learn about the quality of operations taking place at each end of the process, one needs to dissociate perspectives but under $\boldsymbol{UNEX^{TM}}$, but ***IPC*** does not disentangle inbound from outbound stretches. Yet, the focus can be put on each end, disjointly, by averaging the $\widehat{\Pi}(Co2Cd,\mathbb{t})$

- either over all destinations $(d \neq Co)$ for each origin (Co), to get a country-to-Europe cursor:

$$\widehat{\boldsymbol{\Pi}}(\boldsymbol{Co2Eu},\mathbb{t}) = \sum_{d\neq Co}\left(\sum_{s\in S(Co2d)}\left[\omega^{*+}_{s|Co2d} \times \widehat{\Pi}_{s,\mathbb{t}}\right]\right), with: \sum_{d\neq Co}\sum_{s\in S(Co2d)} \omega^{*+}_{s|Co2d} = 1, and\ \forall Co, \tag{3.1}$$

- or over all origins $(o \neq Cd)$ for each destination (Cd), offering a Europe-to-country outlook:

$$\widehat{\boldsymbol{\Pi}}(\boldsymbol{Eu2Cd},\mathbb{t}) = \sum_{o\neq Cd}\left(\sum_{s\in S(o2Cd)}\left[\omega^{+*}_{s|o2Cd} \times \widehat{\Pi}_{s,\mathbb{t}}\right]\right), with: \sum_{o\neq Cd}\sum_{s\in S(o2Cd)} \omega^{+*}_{s|o2Cd} = 1, and\ \forall Cd. \tag{3.2}$$

Indeed, formula (3.1)/alternatively, (3.2)/ rates the efficiency of the outbound /alternatively, inbound/ processes operated by the postal player in the country of origin /alternatively, destination/. To be fair, special caution is in order when referring to such gradings for ranking actors involved because the symbol $\boldsymbol{Eu}$ stands for different truncated sets from which the country under evaluation is excluded (since within-country domestic mail is not surveyed by $\boldsymbol{UNEX^{TM}}$). Calculations of the relevant weights - $\omega^{*+}_{s|Co2d}$ and $\omega^{+*}_{s|o2Cd}$ - are detailed in Table 3. These ensue from weighting the $\omega^{**}_{s|o2d}$ by the ratios of stream-specific volume shares to *total shares in outbound- and inbound-volumes*, $\mathbb{v}_{Co2Eu}$ and $\mathbb{v}_{Eu2Cd}$, respectively.

Table 3. Weightings specific to disjoined *KPIs*.

Conditional weighting for the aggregate outbound-specific *Co2Eu-KPI*:

Using subscript $Co2d$ to identify the path with specific origin in Co to any other destination in $d \neq Co$,

$$\omega^{*+}_{s|Co2d} = \omega_{s|Co2d} \times (\mathbb{v}_{Co2d}/\mathbb{V}_{Co2Eu}), \text{for } s \in S(Co2d) \text{ and } \forall d \neq Co, \quad (\text{T. } 3.1)$$

$\mathbb{V}_{Co2Eu} = \sum_{d \neq Co} \mathbb{v}_{Co2d}$ the share of total real mail volume flowing from Co to any other country within $\boldsymbol{Eu}$.

Sum constraint holding for every outbound country ($\boldsymbol{Co}$):

$$\sum_{d \neq Co} \sum_{s \in S(Co2d)} \omega^{*+}_{s|Co2d} = \sum_{d \neq Co} \left(\frac{\mathbb{v}_{Co2d}}{\mathbb{V}_{Co2Eu}}\right) \cdot \left[\sum_{s \in S(Co2d)} \omega_{s|Co2d}\right] = 1.$$

Conditional weighting for the aggregate inbound-specific *Eu2Cd-KPI*:

Using subscript $o2Cd$ to identify the path with specific destination in Cd from any other origin in o$\neq Cd$,

$$\omega^{+*}_{s|o2Cd} = \omega_{s|o2Cd} \times (\mathbb{v}_{o2Cd}/\mathbb{V}_{Eu2Cd}), \text{ for } s \in S(o2Cd) \text{ and } \forall o \neq Cd, \quad (\text{T. } 3.2)$$

$\mathbb{V}_{Eu2Cd} = \sum_{o \neq Cd} \mathbb{v}_{o2Cd}$, the share of total real mail volume flowing to Cd from any other country within $\boldsymbol{Eu}$.

Sum constraint holding for every inbound country ($\boldsymbol{Cd}$):

$$\sum_{o \neq Cd} \sum_{s \in S(o2Cd)} \omega^{+*}_{s|o2Cd} = \sum_{o \neq Cd} \left(\frac{\mathbb{v}_{o2Cd}}{\mathbb{V}_{Eu2Cd}}\right) \cdot \left[\sum_{s \in S(o2Cd)} \omega_{s|o2Cd}\right] = 1.$$

3.2.3. Consistency of *KPIs*

The superordinate *Europe-to-Europe KPI*, $\widehat{\boldsymbol{\Pi}}(\boldsymbol{Eu2Eu}, \mathbb{t})$**,** should also be arrived at through weighted averages of either the outbound, or the inbound *KPIs*. Such is the case, if one averages the $\widehat{\boldsymbol{\Pi}}(\boldsymbol{o2Eu}, \mathbb{t})$-*KPIs*, over all o-origins, each with a weight equal to its share in the total real mail volume sent from it to the rest of Europe $(\mathbb{V}_{o2Eu})$:

$$\text{using (3.1)} \Longrightarrow \sum_{o} \left[V_{o2Eu} \times \widehat{\boldsymbol{\Pi}}(\boldsymbol{o2Eu}, \mathbb{t})\right] = \sum_{o} \left[\sum_{d \neq o} \sum_{s \in S(o2d)} \left[\{\omega_{s|o2d} \times \mathbb{v}_{o2d}\} \times \widehat{\Pi}_{s,\mathbb{t}}\right]\right] = \widehat{\boldsymbol{\Pi}}(\boldsymbol{Eu2Eu}, \mathbb{t}).$$

Similarly, one also gets back to it when one averages the $\widehat{\boldsymbol{\Pi}}(\boldsymbol{Eu2d}, \mathbb{t})$-KPIs, over all d-destinations, each with a weight equal to its share in the total real mail volume sent to it from the rest of Europe $(\mathbb{V}_{Eu2d})$:

$$\text{using (3.2)} \Longrightarrow \sum_{d} \left[\mathbb{V}_{Eu2d} \times \widehat{\boldsymbol{\Pi}}(\boldsymbol{Eu2d}, \mathbb{t})\right] = \sum_{d} \left[\sum_{o \neq d} \sum_{s \in S(o2d)} \left[\{\omega_{s|o2d} \times \mathbb{v}_{o2d}\} \times \widehat{\Pi}_{s,\mathbb{t}}\right]\right] = \widehat{\boldsymbol{\Pi}}(\boldsymbol{Eu2Eu}, \mathbb{t}).$$

Ultimately, substituting (1) into (2), it gets validated as the volume-weighted average of the stream-specific ones:

$$\widehat{\boldsymbol{\Pi}}(\boldsymbol{Eu2Eu}, \mathbb{t}) = \sum_{o2d} \mathbb{v}_{o2d} \times \left[\sum_{s \in S(o2d)} \omega_{s|o2d} \times \widehat{\Pi}_{s,\mathbb{t}}\right] = \sum_{o2d} \mathbb{v}_{o2d} \times \widehat{\boldsymbol{\Pi}}(\boldsymbol{o2d}, \mathbb{t}).$$

3.3. Shortage of coverage calling for model-based extrapolations

From 2021 on, the address labelling $(f = Ad)$ revealed to be a non-discriminant factor[12]. Hence, for 2023, the *TNS* reduced to: 1,446,336. Yet, the *actual* number of relevant strata (ANS) - i.e., those for which $\omega^{**}_{s|o2d} > 0$ - amounted to:

$$ANS = \sum_{Co} \sum_{Cd \neq Co} |S(Co2Cd)| = 333{,}792.$$

Although this ANS lies much below the TNS, assessing all these hundreds of thousands of strata, albeit each by one single *test-item*, is just impossible, not only because it would be too costly, but even more importantly, because on some rural routes the number of test-items would overflow the real volumes exchanged. In 2023, with a sample of size: $n = 105{,}889$ valid test-items, the number of strata effectively surveyed (SS) hardly exceeded fifty thousand: $SS = 51{,}869$, accounting for only 15.54% $(ss = SS/ANS)$ of the universe. Letting CS denote the subset of checked strata, the real *universe coverage rate* (UCR) is determined by adding up their marginal weights:

$$UCR = \sum_{o2d} \sum_{s \in S(o2d)} \omega^{**}_{s|o2d} = 70.73\%.$$

While that rate stands much higher than the percentage of surveyed strata, it is grossly insufficient to give any hope to draw unbiased estimates of $\widehat{\boldsymbol{\Pi}}$-punctuality measurements from the $\boldsymbol{UNEX^{TM}}$ annual study. This is why we could not just rely on the observed sample proportions of items delivered within the $\mathbb{t}$-day deadline to guess the $\widehat{\Pi}_{s,\mathbb{t}}$, since more than 84% - i.e., $1 - ss$ - of these estimates are missing. Consequently, we have built and calibrated econometric models to predict all $\widehat{\Pi}_{s,\mathbb{t}}$ and therefrom, get the aggregate *KPIs*:

$\widehat{\boldsymbol{\Pi}}(\boldsymbol{O2D}, \mathbb{t})$, for $O2D \in \{Co2Cd, Co2Eu, Eu2Cd\}$, defined here above by: (1), (2), (3.1) and (3.2).

Such an econometric inference process *"borrows strength"* across the whole data set through formalized explicit mathematical relationships between measurements and their determinants, links that can be extrapolated to the non-covered strata. As to the others, all sparingly monitored[13], such *indirect model-based estimates* are expected to be more reliable than those derived by direct weighting of corresponding sample proportions[14].

3.4. Tackling the uncertainty hanging over the weights

In § 3.1.2 here above, we confessed that the so-called *"standard"* weighting basis - despite it is designed in compliance with *CEN*'s instructions - relies on the unrealistic hypothesis of independence between mail features because ***IPC*** does not have the capability to gather detailed data on the repartition of *Co2Cd* real mail volumes exchanged down to the strata-level, nor do they have the means to collect them. Hence, we must acknowledge that the information available on the postal universe is far from complete and realize that such imperfect knowledge entails an increase in the inaccuracy of *KPIs'* valuation. The imprecision due to sampling gets indeed compounded by the lack of dependability of the stratum weights, which may also cause biases. Per country-to-country stream, an infinity of sets of non-negative sum-constrained (adding up to 1) real numbers are eligible as weights' vectors:

$$\boldsymbol{\mathcal{W}_{Co2Cd}} = \left[w_{s|Co2Cd} \geq 0 | s \in S(Co2Cd)\right], \text{ with: } \sum_{s \in S(Co2Cd)} w_{s|Co2Cd} = 1,$$

provided each of their mode-specific $m(f|C)$ subsets - i.e., for which: $I^{s|Co2Cd}_{m(f|C)} = 1$ - jointly match each of the related real mail weights: $\Omega_{m(f|C)}$.

---

[12] Over time, the facsimile of human *handwriting* on test-envelops has become as readable by sorters as the machine-typed addresses.

[13] On average, about 2 items were tested per surveyed stratum: n/SS.

[14] Cf. literature on *"small area / small domain estimation"*: Rao and Molina (2015), Sugasawa and Kubokawa (2023).

Out of all feasible vectors, two extremes are of interest:

- $\underline{\mathcal{W}}_{Co2Cd}$ minimizing the value of the point-estimate of the *KPI*, thus yielding its lower bound: $\underline{\hat{\hat{\Pi}}}(Co2Cd, \mathbb{t})$,
- $\overline{\mathcal{W}}_{Co2Cd}$ maximizing the value of the point-estimate of the *KPI*, thus yielding its upper bound: $\overline{\hat{\hat{\Pi}}}(Co2Cd, \mathbb{t})$.

The double hat-accent above the Π-symbol distinguishes here the extremes from the point estimate (single hat) derived using the *standard weighting basis*. This notation is meant to signal that *KPI*-estimates are subject to two sources of errors: sampling and ex-post weighting. Table 4 specifies the mathematical ***linear programs*** to be solved to obtain these bounds.

Table 4. Determination of the variability of the point-estimate of the *Co2Cd-KPI*.

I. CRITERION
$\hat{\hat{\Pi}}(\boldsymbol{Co2Cd}, \mathbb{t} \mid \boldsymbol{\mathcal{W}_{Co2Cd}}) = \sum_{s \in S(Co2Cd)} \boldsymbol{w_{s\mid Co2Cd}} \times \hat{\Pi}_{s,\mathbb{t}}$, where: $\boldsymbol{\mathcal{W}_{Co2Cd}} = [\boldsymbol{w_{s\mid Co2Cd}} \mid s \in S(Co2Cd)]$ is the vector of unknown strata-specific weights.
II. CONSTRAINSTS
II.1. Non-negativity: $\boldsymbol{w_{s\mid Co2Cd}} \geq \mathbf{0}, \forall s \in S(Co2Cd)$.
II.2. Matching conditional mode-specific real mail weights: $\sum_{s \in S(Co2Cd)} \boldsymbol{w_{s\mid Co2Cd}} \times I^{s\mid Co2Cd}_{m(f\mid C)} = \boldsymbol{\Omega_{m(f\mid C)}}$, for $C \in \{Co, Cd\}, f\mid Co \in \{Uo, Sw, Ad, Fk, Wd, Pl\}$ and $f\mid Cd \in \{Ud\}$.
II.3. Overall consistency sum-constraint: $\sum_{s \in S(Co2Cd)} \boldsymbol{w_{s\mid Co2Cd}} = \mathbf{1}$.
III. OBJECTIVES: SEARCH FOR BOUNDS ON *KPIs*' POINT-ESTIMATES
Lower bound: $\underline{\hat{\hat{\Pi}}}(Co2Cd, \mathbb{t}) = \hat{\hat{\Pi}}(\boldsymbol{Co2Cd}, \mathbb{t} \mid \underline{\boldsymbol{\mathcal{W}}}_{\boldsymbol{Co2Cd}}) = \underset{\{w_{s\mid Co2Cd} \mid s \in S(Co2Cd)\}}{MIN} \hat{\hat{\Pi}}(Co2Cd, \mathbb{t} \mid \mathcal{W}_{Co2Cd})$. Upper bound: $\overline{\hat{\hat{\Pi}}}(Co2Cd, \mathbb{t}) = \hat{\hat{\Pi}}(Co2Cd, \mathbb{t} \mid \overline{\boldsymbol{\mathcal{W}}}_{\boldsymbol{Co2Cd}}) = \underset{\{w_{s\mid Co2Cd} \mid s \in S(Co2Cd)\}}{MAX} \hat{\hat{\Pi}}(Co2Cd, \mathbb{t} \mid \boldsymbol{\mathcal{W}_{Co2Cd}})$.
Solution-vectors
$\underline{\boldsymbol{\mathcal{W}}}_{\boldsymbol{Co2Cd}} = [\underline{\boldsymbol{w}}_{\boldsymbol{s\mid Co2Cd}} \mid \boldsymbol{s \in S(Co2Cd)}] \equiv \underset{\{w_{s\mid Co2Cd} \mid s \in S(Co2Cd)\}}{arg\ min} \hat{\hat{\Pi}}(Co2Cd, \mathbb{t} \mid \mathcal{W}_{Co2Cd})$. $\overline{\mathcal{W}}_{Co2Cd} = [\overline{w}_{s\mid Co2Cd} \mid s \in S(Co2Cd)] \equiv \underset{\{w_{s\mid Co2Cd} \mid s \in S(Co2Cd)\}}{arg\ max} \hat{\hat{\Pi}}(Co2Cd, \mathbb{t} \mid \boldsymbol{\mathcal{W}_{Co2Cd}})$.

Per construction, they delimit the range of values which the point-estimate of the *KPI* can take:

$$\left[\underline{\hat{\hat{\Pi}}}(Co2Cd, \mathbb{t}) = \sum_{s \in S(Co2Cd)} \underline{w}_{s\mid Co2Cd} \times \hat{\Pi}_{s,\mathbb{t}}\right] \leq \hat{\Pi}(Co2Cd, \mathbb{t}) \leq \left[\overline{\hat{\hat{\Pi}}}(Co2Cd, \mathbb{t}) = \sum_{s \in S(Co2Cd)} \overline{w}_{s\mid Co2Cd} \times \hat{\Pi}_{s,\mathbb{t}}\right].$$

Since the summative *KPIs* are mere weighted averages - with known weights: $\mathbb{V}_{Co2Eu}$, $\mathbb{V}_{Eu2Cd}$ and $\mathbb{v}_{Co2Cd}$ – of the *Co2Cd-KPIs*, maximizing/minimizing the latter necessarily leads to maximizing/minimizing the former which result from their aggregation. We programmed the search for these extrema, for each of the 710 *Co2Cd*-streams surveyed in 2023, with the help of the *OR* (*Operations Research*) package of *Statistical Analysis Software* (*SAS*).

The computer code listed in annex **A.1** solves:

$$MAX/MIN\ \widehat{\widehat{\Pi}}(o2d,\mathbb{t}|\ \boldsymbol{\mathcal{W}_{Co2Cd}}) \Rightarrow MAX/MIN \begin{cases} \widehat{\widehat{\Pi}}(Co2Eu,\mathbb{t}|\ \boldsymbol{\mathcal{W}_{Co2d}}) = \sum_{d\neq Co}\left(\frac{\mathbb{v}_{Co2d}}{\mathbb{V}_{Co2Eu}}\right).\widehat{\widehat{\Pi}}(Co2d,\mathbb{t}|\ \boldsymbol{\mathcal{W}_{Co2d}}) \\ \widehat{\widehat{\Pi}}(Eu2Cd,\mathbb{t}|\ \boldsymbol{\mathcal{W}_{o2Cd}}) = \sum_{o\neq Cd}\left(\frac{\mathbb{v}_{o2Cd}}{\mathbb{V}_{Eu2Cd}}\right).\widehat{\widehat{\Pi}}(o2Cd,\mathbb{t}|\ \boldsymbol{\mathcal{W}}_{o2Cd}) \\ \widehat{\widehat{\Pi}}(Eu2Eu,\mathbb{t}|\ \boldsymbol{\mathcal{W}}) = \sum_{o2d}\mathbb{v}_{o2d}\times\widehat{\widehat{\Pi}}(o2d,\mathbb{t}|\ \boldsymbol{\mathcal{W}_{o2d}}). \end{cases}$$

Indeed, simple linear combinations of extrema of $Co2Cd$-*KPIs* yield corresponding extrema of aggregate ones:

$$\underline{\widehat{\widehat{\Pi}}}(\boldsymbol{Co2Eu},\mathbb{t}) = \sum_{d\neq Co}\left(\frac{\mathbb{v}_{\boldsymbol{Co2d}}}{\mathbb{V}_{\boldsymbol{Co2Eu}}}\right).\underline{\widehat{\widehat{\Pi}}}(Co2d,\mathbb{t}) \leftrightarrow \overline{\widehat{\widehat{\Pi}}}(\boldsymbol{Co2Eu},\mathbb{t}) = \sum_{d\neq Co}\left(\frac{\mathbb{v}_{\boldsymbol{Co2d}}}{\mathbb{V}_{\boldsymbol{Co2Eu}}}\right).\overline{\widehat{\widehat{\Pi}}}(Co2d,\mathbb{t}),$$

$$\underline{\widehat{\widehat{\Pi}}}(\boldsymbol{Eu2Cd},\mathbb{t}) = \sum_{o\neq Cd}\left(\frac{\mathbb{v}_{\boldsymbol{o2Cd}}}{\mathbb{V}_{\boldsymbol{Eu2Cd}}}\right).\underline{\widehat{\widehat{\Pi}}}(o2Cd,\mathbb{t}) \leftrightarrow \overline{\widehat{\widehat{\Pi}}}(\boldsymbol{Eu2Cd},\mathbb{t}) = \sum_{o\neq Cd}\left(\frac{\mathbb{v}_{\boldsymbol{o2Cd}}}{\mathbb{V}_{\boldsymbol{Eu2Cd}}}\right).\overline{\widehat{\widehat{\Pi}}}(o2Cd,\mathbb{t}),$$

$$\underline{\widehat{\widehat{\Pi}}}(\boldsymbol{Eu2Eu},\mathbb{t}) = \sum_{o2d}\mathbb{v}_{o2d}\times\underline{\widehat{\widehat{\Pi}}}(o2d,\mathbb{t}) \leftrightarrow \overline{\widehat{\widehat{\Pi}}}(\boldsymbol{Eu2Eu},\mathbb{t}) = \sum_{o2d}\mathbb{v}_{o2d}\times\overline{\widehat{\widehat{\Pi}}}(o2d,\mathbb{t}).$$

4. Mathematical modeling of delivery punctuality

All KPIs of interest (introduced here above in § 3.2.1 and 3.2.2) rest on estimated stratum-level probabilities of delivery within fixed deadlines. To specify these mathematically, one must concentrate on the *delivery time*, operationalized by the number of working days it takes Posts to carry an item - singled out by subscript: i - from its shipping point to its destination. That duration must be treated as a *discrete random variable*, taking integer values greater or equal to one day. Hence, we label it with the tilde $\sim$ accent: $\widetilde{\mathbb{T}}_i$. Characterizing its distribution and fitting it to observed time records will enable us to calculate the KPI inputs, since the probability that item i be delivered to its addressee within a certain $\mathbb{t}$-day deadline is the probability its delivery time is at most equal to $\mathbb{t}$ days:

$$\Pi_{i,\mathbb{t}} = P\left(\widetilde{\mathbb{T}}_i \leq \mathbb{t}\right), \qquad \text{for } \mathbb{t} \in \{1, 2, \dots\}. \tag{4.1}$$

Expression (4.1) corresponds to the *cumulative distribution function* (cdf) of $\widetilde{\mathbb{T}}_i$. Most often the cdf of a discrete variable is determined from its *probability mass function* (pmf) by simple addition:

$$\Pi_{i,\mathbb{t}} = \sum_{t=1}^{t=\mathbb{t}} \pi_{i,t}\,, \text{where: } \pi_{i,t} = P\left(\widetilde{\mathbb{T}}_i = t\right). \tag{4.2}$$

In the present context where about 85% of the $\Pi_{s,\mathbb{t}}$ are to be extrapolated ($1 - ss = 84.46\%$, cf. *3.3*) from records of delivery times, $\widetilde{\mathbb{T}}_i$ is to be delt with as the postal *response* to the factors formalized in the second column of Table 1. To do so, a *predictor function*, linearly combining the effects of these explanatory variables must be specified. In the jargon of certain disciplines such as actuarial sciences, it is also known as the linear *score* (Denuit et al., 2019, pp.100-101). In the sequel, this comprehensive predictor – resulting from all specificities that characterize the item and its route, weighted by parameters reflecting their differential effects on the postal logistic effectiveness – will be interpreted as the *QoS* and labelled: Q_i.

Here, given that all *predictors are the nominal classification variables* listed in Table 1, this *latent construct*, Q_i, is not directly measurable but defined by:

$$Q_i = \mathbf{B} + \left(\sum_{f \in \mathcal{F}} \left[\sum_{m(f) \in \{\boldsymbol{M^f} | m(f) \neq b(f)\}} \delta^f_{m(f)} . x^f_{i,m(f)} \right] \right) + \left(\sum_{o \in \mathcal{O}} v^O_o . Z^O_{i,o} + \sum_{d \in \mathcal{D}} v^D_d . Z^D_{i,d} \right). \quad (5)$$

where:

- the superscript f identifies the f^{th} predictor, while the second subscript m points the mode of f characterizing item i, and $\boldsymbol{M^f}$ is the set of alternative modes that f may take,

- the variable $x^f_{i,m(f)}$ is the $1/0$ binary dummy indicating whether, or not, $m(f)$ is indeed the mode characterizing the i^{th} item with respect to f, and its coefficient differentiates the mode effect from an arbitrarily chosen benchmark denoted $b(f)$:

$$\delta^f_{m(f)} = \beta^f_{m(f)} - \beta^f_{b(f)} \quad (5.1)$$

where $\beta^f_{m(f)}$ is the underlying regression coefficient, and $\beta^f_{b(f)}$ is the baseline parameter,

- consequently, **B** stands for the reference effectiveness level, equal to the sum of the baseline parameters, unidentifiable separately:

$$\mathbf{B} = \sum_{f \in \mathcal{F}} \beta^f_{b(f)} \quad (5.2)$$

- the inner sum - within square brackets in the second term **-** defines, without loss of generality, the effect of the f-predictor on i's handling,

- $Z^O_{i,o}$ is the $1/0$ binary dummy indicating whether, or not, the i^{th} item was sent from the o^{th} postal area, while v^O_o is the random component reflecting local idiosyncrasies of outbound logistics within that zone,

- $Z^D_{i,d}$ is the $1/0$ binary dummy indicating whether, or not, the i^{th} item was sent to the d^{th} postal area, while v^D_d is the random component reflecting local idiosyncrasies of inbound logistics within that zone.

The random components, v^O_o and v^D_d, are meant to encompass spatial heterogeneity in logistics across on the one hand, origin areas and on the other, destination vicinities. They are assumed to be zero-mean mutually independent, normally distributed, random variables, with respective standard deviations:

$$Stdv[v^O_o] = \varsigma^0, \quad (5.3)$$

and

$$Stdv[v^D_d] = \varsigma^D. \quad (5.4)$$

In accordance with the *KISS*-principle - as reinterpreted by Arnold Zellner (2002) to mean: ***k**eeping **i**t* (i.e., the model) ***s**ophisticatedly **s**imple* -, we defined these zones by the country-urbanization pairings:

$$o \in \{\boldsymbol{M^{Co}} \cap \boldsymbol{M^{Uo}}\} \Rightarrow x^{Co}_{i,m(Co)} . x^{Uo}_{i,m(Uo)} \Rightarrow Z^O_{i,o}, \quad (5.5)$$

and

$$d \in \{\boldsymbol{M^{Cd}} \cap \boldsymbol{M^{Ud}}\} \Rightarrow x^{Cd}_{i,m(Cd)} . x^{Ud}_{i,m(Ud)} \Rightarrow Z^D_{i,d}. \quad (5.6)$$

The resulting *random intercepts* suffice to capture the essentials of the spatial heterogeneity at both the outbound- and inbound-ends. They are introduced in the model to reflect uncontrollable events occurring during the first- and last-miles of the postal journey, close to the interfaces of Posts with their customers: between senders and the outbound logistics on one side, and between inbound logistics and the final addressee on the other side[15].

[15] *"Territory design for last-mile delivery faces the challenge that regions have to be determined without deterministic knowledge on the varying sets of customers that have to be serviced each day"* (Boysen et al., 2021, p. 10). Comparable uncertainties in demand dispersion must be coped with for the first mile.

The number of these intercepts remains manageable[16]. Modeling finer mappings of the postal logistics space, up to considering panelists as clusters, would cause serious problems at the parametrization stage: slow convergence (prohibitively long running time) of estimation heuristics, infeasibility or divergence, computer memory shortage (Kiernan et al., 2012).

Predicting values of either the $\pi_{i,t}$, or the $\Pi_{i,\mathbb{t}}$, from Q_i, further demands to shape and calibrate a consistent, hence non-linear, *econometric link* between the *predictor function* and the *pmf (probability mass function)*, or the *cpf (cumulative probability function)* of $\widetilde{\mathbb{T}}_i$. The next two subsections explain how such links can be established: the first applies *count regression*, while the second relates to *ordinal regression*.

4.1. Specification via the probability mass function

Starting from the *pmf* of $\widetilde{\mathbb{T}}_i$ amounts to regarding delivery time as *a discrete nonnegative integer*. The most popular approach to the analysis of such noncontinuous variates is known as *count* regression designed to predict *responses* resulting from the count of the number of occurrences $(\widetilde{\mathbb{N}})$ of a specific event, happening randomly and independently at a constant rate, within a limited time frame, and/or in a restricted spatial area and/or under peculiar circumstances. The *pmf* on which *count* regression was built is the *Poisson* law, limit of the *binomial* distribution,

$$P\left(\widetilde{\mathbb{N}}_i = \mathbb{m}\right) = [\mu_i^{\mathbb{m}}.\, e^{-\mu_i}]/\mathbb{m}!\,, \text{for } \mathbb{m} = 0, 1, 2, \ldots \tag{6}$$

where $\mu_i > 0$, which depends on the context i, is both the conditional expectation and variance of $\widetilde{\mathbb{N}}_i$.

4.1.1. Negative binomial probability mass function

In (6), μ_i shapes the central tendency, as well as the dispersion of the $\widetilde{\mathbb{N}}_i$*-pmf*, which thus lacks plasticity. Moreover, μ_i is assumed to be completely determined by its presupposed predictors. Yet, introducing a nonnegative random component reflecting the pervasive uncertainty $(\tilde{u}_i)$ can only add flexibility and realism. Therefore, the *Poisson* model was generalized to account for possible *heterogeneity* in the stochastic process generating $\widetilde{\mathbb{N}}_i$ by substituting for μ_i into (6): $\tilde{\mu}_i = \mu_i.\tilde{u}_i$, to yield:

$$P\left(\widetilde{\mathbb{N}}_i = \mathbb{m} | \tilde{u}_i = u\right) = \left[(\mu_i.u)^{\mathbb{m}}.\, e^{-(\mu_i.u)}\right]/\mathbb{m}! \tag{7.1}$$

Assuming further that the random multiplier $\tilde{u}_i$ is a *gamma distributed noise*, whose probability density function is:

$$g(\tilde{u}_i = u) = \left(\frac{\gamma^\gamma}{\Gamma(\gamma)}\right).\left[u^{(\gamma-1)}.\, e^{-\gamma.u}\right], \text{for: } u \in [0, \infty[, \text{ and where } \Gamma(.) \text{ is the generalized factorial,} \tag{7.2}$$

ensures that $E[\tilde{\mu}_i] = \mu_i$, because: $E[\tilde{u}_i] = 1$. So, $\tilde{u}_i$ either amplifies (if $u > 1$) or dampens (if $u < 1$) the effects of the fixed predictors encapsulated by μ_i, according to its variance which is the inverse of the single parameter characterizing its distribution: $V[\tilde{u}_i] = \gamma^{-1}$. Function (7.2) was chosen because it is the *natural conjugate prior* of (7.1), because the very fact that their *kernels* take the same form eases their mixing (Hilbe, 2011, §8.2.1: 188-193):

$$P\left(\widetilde{\mathbb{N}}_i = \mathbb{m}\right) = \int_{u=0}^{u=\infty} P\left(\widetilde{\mathbb{N}}_i = \mathbb{m} | \tilde{u}_i = u\right).g(\tilde{u}_i = u).du = \left(\frac{\gamma^\gamma}{\Gamma(\gamma)}.\frac{\mu_i^{\mathbb{m}}}{\mathbb{m}!}\right).\int_{u=0}^{u=\infty} u^{(\mathbb{m}+\gamma-1)}.\, e^{-(\mu_i+\gamma).u}.du,$$

$$\text{or } P\left(\widetilde{\mathbb{N}}_i = \mathbb{m}\right) = \left(\frac{\gamma^\gamma}{\Gamma(\gamma)}.\frac{\mu_i^{\mathbb{m}}}{\mathbb{m}!}\right).\left(\frac{\Gamma(\mathbb{m}+\gamma)}{(\mu_i+\gamma)^{(\mathbb{m}+\gamma)}}\right).$$

[16] In total, $184 = 2 \times (3 \times 31 - 1)$, for in *Cyprus* (CY), $m(Uo) \neq Ca$ and $m(Ud) \neq Ca$, *Nicosia* being classified as a *key city*.

Ultimately, letting $\phi = \gamma^{-1}$ leads to the ***negative binomial*** (NB) distribution (ibidem, equation (8.6), p. 189):

$$P(\widetilde{\mathbb{N}}_i = \mathbb{m}) = \frac{\Gamma(\mathbb{m} + (1/\phi))}{\Gamma(1/\phi).\Gamma(\mathbb{m}+1)} \cdot \wp_i^{(1/\phi)}.(1-\wp_i)^{\mathbb{m}}, \text{with: } \wp_i = \frac{1}{1+(\phi.\mu_i)}, \qquad \text{for } \mathbb{m} = 0, 1, 2, \ldots \tag{8}$$

such that:

$$V[\widetilde{\mathbb{N}}_i] = \mu_i + (\phi.\mu_i^2), \tag{9}$$

which explains why ϕ has been called the *overdispersion parameter* and shows that the *Poisson pmf* is the equi-dispersion limit of the ***NB**-pmf* since:

$$\lim_{\phi \to 0} V[\widetilde{\mathbb{N}}_i] = \mu_i.$$

Then, it comes as no surprise that such a versatile model was fitted to such various count statistics as the number of units of a branded product item bought on a shopping trip (Ehrenberg, 1959), vehicle accidents on segments of Florida state road 50 over three years (Abdel-Aty and Radwan, 2000), disease biomarkers of patients (Yirga et al, 2020), bat calls collected using acoustic detectors for 2–5 nights across 20 sites (Stoklosa et al., 2022), … Other cases in very diverse universes have been reviewed by Winkelmann (2008, Chapter 9: 251-298). The ***NB** pmf* was also employed to analyse measurements of process duration, for example to depict the lags in the upshots of investments (Solow, 1960; Bultez and Naert, 1979) by the parsimonious generalization of the geometric distribution through the *Pascal pmf* (special case of the *NB*, when γ is an integer), study the effects on workers' absenteeism (measured by the number of non-condonable workdays of absence over a year) of remuneration contracts and other covariates such as gender, sick-pay grade, age, … (Barmby et al., 2001), explain the length-of-stay in hospitals of a specific diagnostic group, an example extensively dealt with by Hilbe, throughout his book (2011, introduced on page 100), weigh the settings (hospital, personal, and visit characteristics) that may impinge on waiting time wasted in hospitals' emergency departments (Cai and Shimizu, 2014), assess the risk factors likely to determine the extent of medical leaves (work disability in days) of victims of motor accidents (Bermúdez et al., 2018).

4.1.2. Right-shifted ***NB***-pmf for nonzero delivery times

From this point on, we consider exclusively ***delivery processes whose duration cannot be less than one day***, which is the case for most international transports. So, to fit the ***NB*** to ***UNEX^{TM}*** records, we must apply ***a one-day right-shift*** to the measured delivery times. It suffices in formula (8) to substitute $(\widetilde{\mathbb{T}}_i - 1)$ for $\widetilde{\mathbb{N}}_i$, and thus $(\mathbb{t}-1)$ for $\mathbb{m}$, which yields:

$$\pi_{i,\mathbb{t}} = P(\widetilde{\mathbb{T}}_i = \mathbb{t}) = \frac{\Gamma(\mathbb{t}-1+\phi^{-1})}{\Gamma(\phi^{-1}).\Gamma(\mathbb{t})} \cdot \wp_i^{(1/\alpha)}.(1-\wp_i)^{\mathbb{t}-1}, \text{with: } \wp_i = \frac{1}{1+(\phi.\mu_i)}, \text{and } \mu_i, \phi > 0, \text{for } \mathbb{t} = 1, 2, \ldots \tag{10.1}$$

The expected value and variance of the delivery time are thus:

$$E(\widetilde{\mathbb{T}}_i) = E(\widetilde{\mathbb{N}}_i + 1) = \mu_i + 1 \tag{10.2}$$

and

$$V(\widetilde{\mathbb{T}}_i) = V(\widetilde{\mathbb{N}}_i + 1) = \mu_i + (\phi.\mu_i^2). \tag{10.3}$$

Formula (10.3) shows how the overdispersion parameter ϕ inflates the variance. Chapter 7 of Hilbe's book (2011, pp. 141-184) is entirely devoted to *real* and *apparent overdispersion*; there, he states upfront that *"few reallife Poisson data sets are truly equidispersed. Overdispersion to some degree is inherent to the vast majority of Poisson data"* (op. cit., p.141). And he lists the main causes of *real overdispersion* often effective in practice: "*positive correlation between responses or … an excess variation between response probabilities or counts"* or *"when the data are clustered"* (ibidem).

Figure **2** - where $\pi_{i,\mathbb{t}}$ is plotted on the vertical axis as a function of $\mathbb{t}$ in abscissa - illustrates how μ_i and ϕ govern the shape, location, and extent of the *NB-pmf* curve.

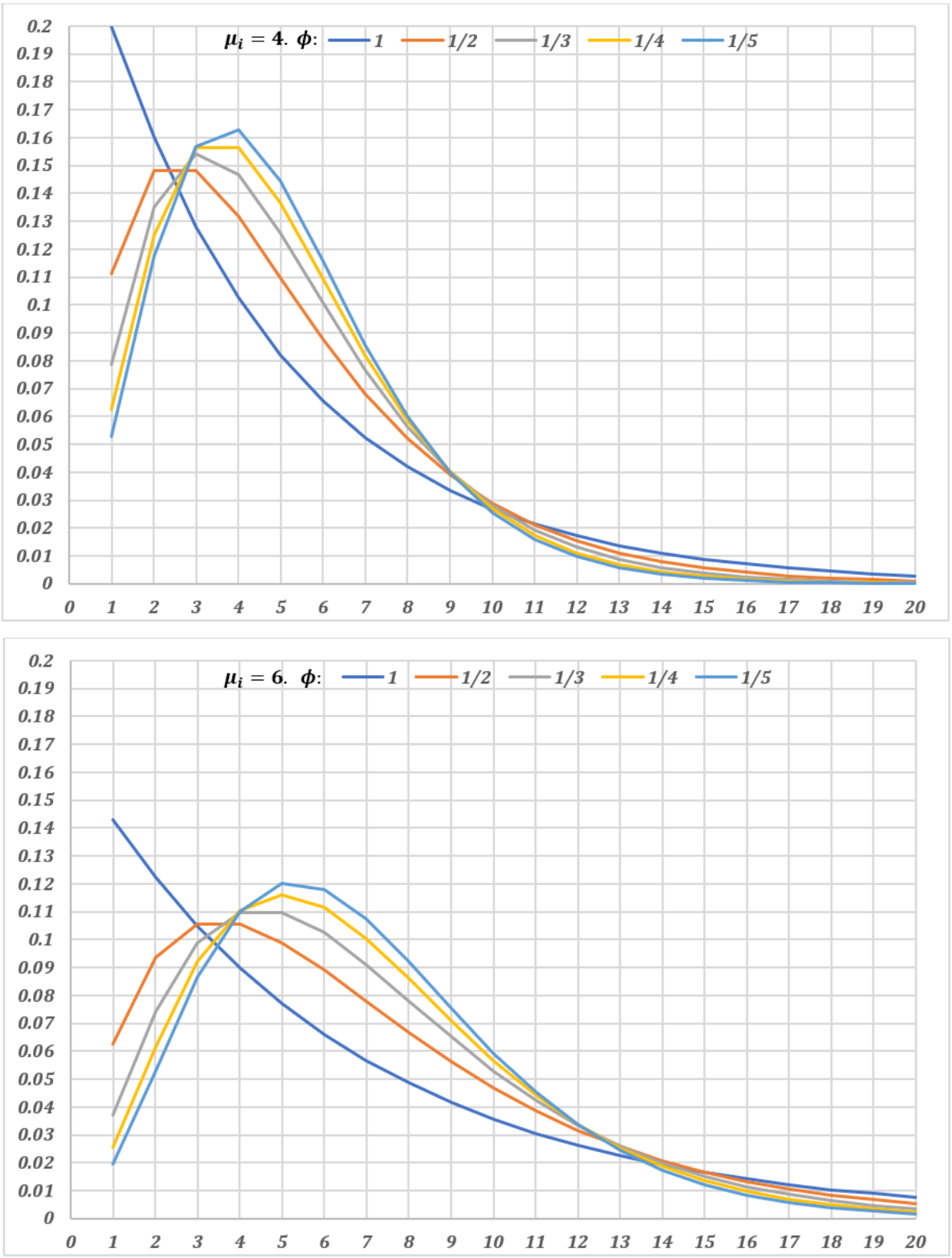


Figure 2. Sensitivity of the ***NB*** *pmf* to its parameters.

4.1.3. Negative binomial regression

Linear predictor functions, such as the one made explicit by (5), are linked to the *pmf* of the count random variable, here defined by (10.1), through its location parameter: i.e., μ_i. Given that

- on the one hand, (5) yields negative values for items difficult to handle, hence with low chances of being delivered with the fixed deadline (small $\pi_{i,\mathbb{t}}$),
- on the other, μ_i can only be positive since $\mathbb{t} \geq 1$,

it is only natural to assume an exponential relationship between the measurement of what we labelled the *QoS* - i.e., Q_i - and the location of the delivery time *pmf*:

$$\mu_i = e^{-Q_i} > 0. \tag{11}$$

The exponential form ensures that μ_i remains positive whatever value Q_i may take. The negative sign appearing in the exponent reflects that the delivery time decreases as the effectiveness of the delivery logistics increases. Also, (11) implies that introducing the heterogeneity component $\tilde{u}_i$ is analogous to adding a zero-mean error term:

$\tilde{\mu}_i = e^{-\tilde{Q}_i}$, where: $\tilde{Q}_i = Q_i + \tilde{\varepsilon}_i, \tilde{\varepsilon}_i = -\ln \tilde{u}_i$, such that $E[\tilde{\varepsilon}_i] \approx -\ln E[\tilde{u}_i] = 0$ (first order approximation). (12)

4.2. Specification via the cumulative distribution function

Chances to meet a $\mathbb{t}$-day target can be determined directly via (4.1). To make their evaluation explicit, let us focus on the *QoS* which should be high enough to keep delivery time within the deadline, which implies that the *QoS* should exceed a certain minimal level, say: $\Theta_{\mathbb{t}}$. Consequently, (4.1) can be developed through the specification of the distribution of the *QoS*:

$$\Pi_{i,\mathbb{t}} = P(\tilde{\mathbb{T}}_i \leq \mathbb{t}) = P(\tilde{Q}_i \geq \Theta_{\mathbb{t}}), \tag{13.1}$$

where the tilde accent put on Q_i indicates that from now on we treat it as a random variable: $Q_i \Longrightarrow \tilde{Q}_i$. Indeed, besides the fixed constituents of the linear predictor function, other numerous, unmanageable elements, each of minor influence, may alter postal efficiency. Hence, their total impact can be modelled through a disturbance term, $\tilde{\varepsilon}_i$, that add noise to the deterministic model (5), thereby acknowledging that predictions of Q_i are uncertain, as in (12):

$$\tilde{Q}_i = Q_i + \tilde{\varepsilon}_i \tag{13.2}$$

On average, the effects of the uncontrollable components, which $\tilde{\varepsilon}_i$ embodies, are assumed to compensate one another so that the probability distribution of $\tilde{\varepsilon}_i$ is symmetric around, and peaking at, zero. So, its expected value (mode and median, as well) is exactly zero. This postulate entails no loss of generality because a non-zero mean would be picked up by the model intercept (i. e. , **B**).

Of course, the shorter the $\mathbb{t}$-day deadline is, the higher the value of the corresponding $\Theta_{\mathbb{t}}$-***threshold*** should be. Practically, measurements of probabilities of delivery by $\mathbb{t} = 1$ up to $\mathbb{t} = \bar{\mathbb{t}}$, day-by-day, require $\bar{\mathbb{t}}$ parameters, since in that case, (13.1) applies for $\mathbb{t} \in \{1,2,3, \dots, \bar{\mathbb{t}}\}$, with: $\Theta_1 > \Theta_2 > \Theta_3 > \cdots > \Theta_{\bar{\mathbb{t}}}$. Thus, such ranking of delivery performances, requires $\bar{\mathbb{t}} + 1$ classes, or ordered levels: $\{1, 2, \dots, \bar{\mathbb{t}}, \mathbb{L}\}$, where $\mathbb{L}$ stands for the last category formed by all late arrivals, i.e., exceeding the $\bar{\mathbb{t}}$-day upper limit. Thus, with $\bar{\bar{\mathbb{t}}}$ denoting the longest observable delivery time, $\mathbb{L} = \{(\bar{\mathbb{t}} + 1) \cup (\bar{\mathbb{t}} + 2) \cup (\bar{\mathbb{t}} + 3) \dots \cup \dots \cup \bar{\bar{\mathbb{t}}}\}$. Figure **3** illustrates all possible events:

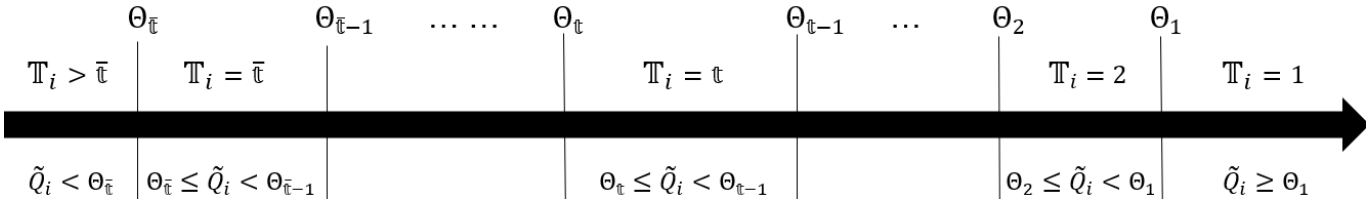


Figure 3. Ordered categories of performances along the quality-of-service continuum.

Substituting (13.2) into (13.1), one obtains: $P(\tilde{\mathbb{T}}_i \leq \mathbb{t}) = P((Q_i + \tilde{\varepsilon}_i) > \Theta_{\mathbb{t}}) = P(\tilde{\varepsilon}_i > (\Theta_{\mathbb{t}} - Q_i))$. This probability can be inferred from the distribution of the standardized disturbance term $\tilde{\varepsilon}_i^*$, by dividing it by its dispersion (σ):

$$P(\tilde{\mathbb{T}}_i \leq \mathbb{t}) = P\left(\frac{\tilde{\varepsilon}_i}{\sigma} > \frac{(\Theta_{\mathbb{t}} - Q_i)}{\sigma}\right) = P(\tilde{\varepsilon}_i^* > (\theta_{\mathbb{t}} - q_i)) = 1 - \mathcal{F}(\theta_{\mathbb{t}} - q_i), \tag{14}$$

where: $\tilde{\varepsilon}_i^* = \tilde{\varepsilon}_i/\sigma$, $\theta_{\mathbb{t}} = \Theta_{\mathbb{t}}/\sigma$, $q_i = Q_i/\sigma$ and $F(.)$ denotes the cumulative density function of $\tilde{\varepsilon}_i^*$. Because of the assumed symmetry around zero of the distribution of $\tilde{\varepsilon}_i^*$, $1 - F(\nu) = F(-\nu)$. Therefore, (14) reduces to:

$$\Pi_{i,\mathbb{t}} = P(\tilde{\mathbb{T}}_i \leq \mathbb{t}) = P(\tilde{Q}_i > \Theta_{\mathbb{t}}) = F(q_{i|\mathbb{t}}), \tag{15.1}$$

with:

$$q_{i|\mathbb{t}} = q_i - \theta_{\mathbb{t}}, \tag{15.2}$$

where according to (5),

$$q_{i|\mathbb{t}} = -\theta_{\mathbb{t}|B} + \left(\sum_{f \in \mathcal{F}} \left[\sum_{m(f) \in \{\boldsymbol{M}^f | m(f) \neq b(f)\}} \delta^f_{m(f)/\sigma}.x^f_{i,m(f)}\right]\right) + \left(\sum_{o \in \mathcal{O}} \upsilon^O_{o/\sigma}.Z^O_{i,o} + \sum_{d \in \mathcal{D}} \upsilon^D_{d/\sigma}.Z^D_{i,d}\right) \tag{15.3}$$

with:

$$\theta_{\mathbb{t}|B} = \theta_{\mathbb{t}} - (\mathbf{B}/\sigma) \tag{15.4}$$

$$\delta^f_{m(f)/\sigma} = \delta^f_{m(f)}/\sigma \tag{15.5}$$

$$\upsilon^O_{o/\sigma} = \upsilon^O_o/\sigma \Longrightarrow Stdv[\upsilon^O_{o/\sigma}] = \varsigma^O/\sigma \tag{15.6}$$

and

$$\upsilon^D_{d/\sigma} = \upsilon^D_d/\sigma \Longrightarrow Stdv\left[\upsilon^D_d\right] = \varsigma^D/\sigma. \tag{15.7}$$

Due to the standardization introduced in (14), all parameters - including the standard deviations of the random components - are deflated (i.e., are scaled in σ-units): thus, their estimates are downsized by the weight (σ) of the unobserved (unexplained) part of the latent variable. Moreover, σ, itself, may vary as potential explanatory factors get added ($\sigma \searrow$) or deleted ($\sigma \nearrow$). Therefore, estimates from different specifications derived from the same sample, and a fortiori from different samples, are not directly comparable (Mood, 2010).

4.2.1. Multinomial probability distribution of delivery times

From the second equality in (13.1) a new *pmf* of delivery times can be derived:

$$\pi_{i,\mathbb{t}} = P(\tilde{\mathbb{T}}_i = \mathbb{t}) = P(\Theta_{\mathbb{t}} < \tilde{Q}_i \leq \Theta_{\mathbb{t}-1}) = P(\tilde{Q}_i \leq \Theta_{\mathbb{t}-1}) - P(\tilde{Q}_i \leq \Theta_{\mathbb{t}})$$

or, according to (15.1):

$$\pi_{i,\mathbb{t}} = \left[1 - F(q_{i|\mathbb{t}-1})\right] - \left[1 - F(q_{i|\mathbb{t}})\right] = F(q_{i|\mathbb{t}}) - F(q_{i|\mathbb{t}-1}) = P(\tilde{Q}_i > \Theta_{\mathbb{t}}) - P(\tilde{Q}_i > \Theta_{\mathbb{t}-1}), \tag{16.1}$$

with for extreme values:

$$\pi_{i,\mathbb{L}} = P(\tilde{\mathbb{T}}_i \geq \bar{\mathbb{t}}) = P(\tilde{Q}_i \leq \Theta_{\bar{\mathbb{t}}}) = 1 - F(q_{i|\bar{\mathbb{t}}}) \tag{16.2}$$

and

$$\pi_{i,1} = P(\tilde{\mathbb{T}}_i = 1) = P(\tilde{Q}_i > \Theta_1) = F(q_{i|1}). \tag{16.3}$$

Equations (16.2) and (16.3) are consistent with (16.1) of which they are indeed special cases:

$$\pi_{i,\mathbb{L}} = P(\Theta_{\mathbb{L}} = -\infty < \tilde{Q}_i \leq \Theta_{\bar{\mathbb{t}}}) = P(\tilde{Q}_i \leq \Theta_{\bar{\mathbb{t}}}) - P(\tilde{Q}_i \leq -\infty) = P(\tilde{Q}_i \leq \Theta_{\bar{\mathbb{t}}}), \text{since: } P(\tilde{Q}_i \leq -\infty) = 0$$

$$\pi_{i,1} = P(\Theta_1 < \tilde{Q}_i \leq \Theta_0 = +\infty) = P(\tilde{Q}_i \leq +\infty) - P(\tilde{Q}_i \leq \Theta_1) = P(\tilde{Q}_i > \Theta_1), \text{since: } P(\tilde{Q}_i \leq +\infty) = 1.$$

4.2.2. Logistically distributed disturbances

The most popular specification of F is the logistic function, Λ, which transposed in the present context, leads to:

$$F(q_{i|\mathbb{t}}) = \Lambda(q_{i|\mathbb{t}}) = 1/[1 + e^{-q_{i|\mathbb{t}}}].$$

Hence, (15.1) takes the explicit form:

$$\Pi_{i,\mathbb{t}} = P(\tilde{\mathbb{T}}_i \leq \mathbb{t}) = P(\tilde{Q}_i > \Theta_{\mathbb{t}}) = 1/[1 + e^{-q_{i|\mathbb{t}}}]. \quad (17)$$

Figure **4** exemplifies how the ***multinomial cumulative logit (MCL)*** model so defined works. More precisely, it shows the implementation of formula (16.1) when (17) holds true. As it should be, chances of meeting deadlines increase with the *QoS*: growth curves of $\Pi_{i,\mathbb{t}}$ and $\Pi_{i,\mathbb{t}-1}$ are S-shaped because of the choice of the logistic *cdf* specified by (17). Moreover, $\Pi_{i,\mathbb{t}} > \Pi_{i,\mathbb{t}-1}$ because $\Theta_{\mathbb{t}-1} > \Theta_{\mathbb{t}}$, since the shorter the deadline, the tougher the challenge, the higher the *QoS* should be. Naturally, the probability that the mail item be received exactly $\mathbb{t}$ days after its sending date, $\pi_{i,\mathbb{t}}$, which results from the difference between the ordinates of the two S-shaped curves, evolves non-monotonically: first it rises, up to a maximum value, beyond which it declines. This non-monotonicity is due to the fact that beyond a certain *QoS*-level, delivery times shorter than the one considered become significantly more likely: $\pi_{i,\mathbb{t}-1} \nearrow \Longrightarrow \pi_{i,\mathbb{t}} \searrow$.

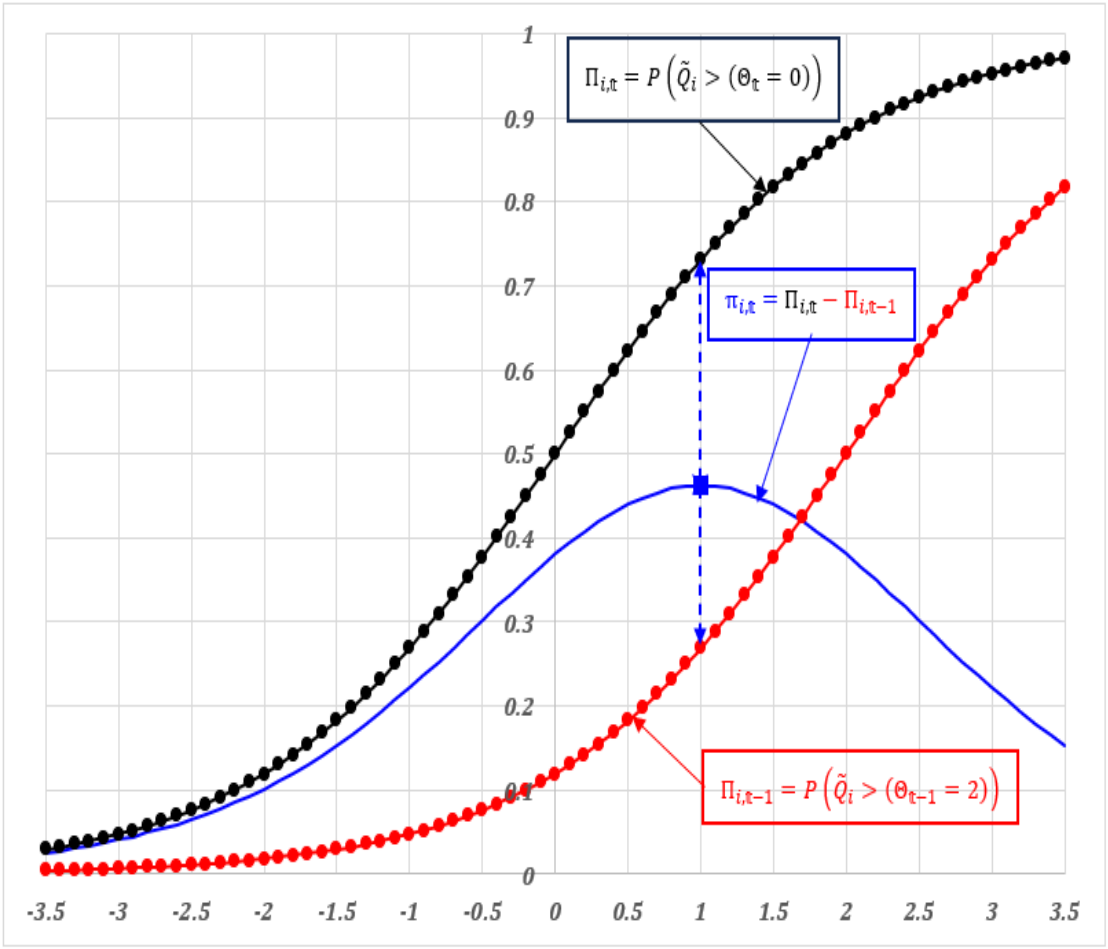


Figure 4. Deduction of probability of delivery time from the logistic modeling of the chances of meeting deadlines, as functions of the *QoS* standardized measure (q_i plotted along the horizontal axis).

4.2.3. Proportional odds property

Equation (17) can be linearized by the *logit* transform, i.e., *the natural logarithm of the odds in favor of reaching the goal*. The odds in favor of the achievement $\left(\widetilde{\mathbb{T}}_i \leq \mathbb{t}\right)$ result from the ratio of the ***chances of being delivered within the deadline:***

$$P\left(\widetilde{\mathbb{T}}_i \leq \mathbb{t}\right) = 1/(1 + e^{-q_{i|\mathbb{t}}}) \tag{18.1}$$

to the ***risk of being delivered later*:**

$$P\left(\widetilde{\mathbb{T}}_i > \mathbb{t}\right) = R_i(\mathbb{t}) = e^{-q_{i|\mathbb{t}}}/(1 + e^{-q_{i|\mathbb{t}}}) = 1/(e^{q_{i|\mathbb{t}}} + 1) = \Lambda\left(-q_{i|\mathbb{t}}\right). \tag{18.2}$$

This ratio simplifies to:

$$\mathbb{O}\left(\widetilde{\mathbb{T}}_i \leq \mathbb{t}\right) = e^{q_{i|\mathbb{t}}}, \tag{18.3}$$

and using (15.2), its logarithm yields:

$$ln\left(\mathbb{O}\left(\widetilde{\mathbb{T}}_i \leq \mathbb{t}\right)\right) = ln(e^{q_{i|\mathbb{t}}}) = q_{i|\mathbb{t}} = q_i - \theta_{\mathbb{t}}. \tag{19}$$

Hence, the *odds ratio* for two items, $i \neq \ell$, entailing different handling efficiency levels, $Q_i \neq Q_\ell$, is independent of the target $\mathbb{t}$ considered, a property called: *"proportional odds"* (McCullagh, 1980, p. 110), since:

$$OR_{i,\ell} = \frac{\mathbb{O}\left(\widetilde{\mathbb{T}}_i \leq \mathbb{t}\right)}{\mathbb{O}\left(\widetilde{\mathbb{T}}_\ell \leq \mathbb{t}\right)} = \frac{P\left(\widetilde{\mathbb{T}}_i \leq \mathbb{t}\right)/P\left(\widetilde{\mathbb{T}}_i > \mathbb{t}\right)}{P\left(\widetilde{\mathbb{T}}_\ell \leq \mathbb{t}\right)/P\left(\widetilde{\mathbb{T}}_\ell > \mathbb{t}\right)} = e^{q_{i|\mathbb{t}}}/e^{q_{\ell|\mathbb{t}}} = e^{(q_i - \theta_{\mathbb{t}})}/e^{(q_\ell - \theta_{\mathbb{t}})} = e^{q_i}/e^{q_\ell} = e^{(q_i - q_\ell)}, \forall \mathbb{t}. \tag{20}$$

In (20), the $\theta_{\mathbb{t}}$ *"cut point"* parameter (McCullagh, ibidem) cancels out and so the conditional subscript $\mathbb{t}$ disappears. This property results from the assumption that the parameters weighting the items' attributes in determining the quality-of-service are independent of the target $\mathbb{t}$. Agresti (2019, p. 177) defends this proportional-odds variant of the cumulative logit model with the logical consistency argument: *"When the model does not fit well, one could consider the more general cumulative model that has separate effects for different cumulative probabilities..."*: i.e.,

the $\delta^f_{m(f)}$ get differentiated into $\delta^f_{m(f)|\mathbb{t}}$.

So, the *"curves for different cumulative probabilities climb or fall at different rates, but then those curves cross at certain predictor values. This is inappropriate because this violates the order that cumulative probabilities must have."* Moreover, this differentiation considerably increases the number of parameters: in the present case, it would multiply it by $\bar{\mathbb{t}}$. As Agresti (op. cit., p. 178) underlines it: *"Even though the model itself* [i.e., its most complete differentiated generalization] *may have less bias, estimates of measures of interest such as odds ratios or category probabilities may be poorer because of the lack of model parsimony. We do not recommend this approach unless the lack of fit of the ordinal model is severe in a practical sense."*

4.2.4. Pervasiveness of logistic regression: predominance of its binomial variant, over its multinomial one

Logistic regression tools have been developed and implemented

- to explain three types of categorical responses: binary ($\mathbb{L} = 2$), multinomial nominal (choices) and multinomial ordinal (the ***MCL***),
- at various degrees of technicality: simple/generalized, one-level/hierarchical, without/with random effects, frequentist/Bayesian,

- in almost all fields:
  - genetics (Wang et al., 2016), epidemiology (Armstrong and Sloan, 1988; Khan et al., 2015),
  - medicine: diagnoses (McCullagh, 1980; Agresti, 2007, pp. 182-184), clinical trials (Zhang et al., 2021), healthcare treatments (Anderson and Philips, 1981, pp. 27-30; Papadopoulos et al., 2021); vaccination uptake (Ross et al., 2022),
  - biology (Shaban and Alkawareek, 2022), zoology (Abts et al., 2018), agriculture-forestry (Uusitalo et al., 2018),
  - education (Peng et al., 2002), sociology (Diekmann et al., 2022), opinion surveys (Dalla Valle et al., 2020), political science (Agresti, 20O2, pp. 502-504; Carrubba et al., 2012; Wolter et al., 2003), public policy (Kim et al., 2013, pp. 171-172, 177, 179), …

Their usefulness has also been demonstrated in management areas: numerous examples can be found in production (White et al.,1999, pp. 9 et seq.), logistics (Castillo et al., 2018), transportation (Farid and Ksaibati, 2021; Pritchard et al., 2021), finance (risk of bankruptcy: Calabrese et al., 2016; chances of mergers and acquisitions: Alam and Lee, 2014), and even anticipations of stock price movement directions (Yang et al., 2022)), in marketing (Guadagni and Little, 1983; de Haan et al, 2015; Bultez et al, 2025) … Up to a point where some have argued in favor of its inclusion within undergraduate *Business Administration* curricula: Brusco (2022), and Hoang and Watson (2022).

On the contrary, implementations of the $\boldsymbol{MCL}$ model have been rare because binary responses are much more frequently studied than multilevel ordinal ones and when these are analyzed authors pool categories up to dichotomizing them. Such mergers of categories, which entail information loss, are acknowledged, and justified more or less explicitly by sample imbalances:

- Dalla Valle et al. (2020, p. 433) collapsed into two degrees the four grades of the Europeans' attitude towards immigrants from poorer countries (*"Allow many/some/a few/none*)[17]: *"The dependent variable is 'immig', indicating whether the respondent would allow immigrants from poorer countries outside Europe, with 'immig = 1' if the respondent is against immigration, and' immig = 0' if the respondent is in favour of immigration. The dependent variable was obtained by dichotomizing the 'ESS' variable 'impcntr'."*
- *"The response modeled was whether the crash was a severe crash involving a fatality or incapacitating injury. The counts of fatal crashes and of incapacitating injury crashes were low. Therefore, they had to be combined."* (Farid and Ksaibati, 2021, p. 228).
- *"The target variable is a simple binary variable of commute satisfaction, where 1 is individuals who are satisfied with their commute, and 0 those who are not. This variable was created from a single 5-point Likert scale question. Respondents who reported being satisfied or very satisfied were assigned as 'satisfied'. While the possibility of treating the variable as ordinal was explored, the response profiles in the different case study regions necessitated the transformation of the variable … As a result, even with good model fit indices, the model was unlikely to correctly predict the individuals in particularly small groups.* (Pritchard et al., 2021, pp. 1000-1001, 1009: Table 5).
- Ross et al (op. cit., p. 5) reduced the seven-point scaling of the vaccination likelihood they studied to two classes: *"All participants that indicated they were likely to receive a vaccine (extremely likely, moderately likely) were grouped together, and those who reported being unlikely (slightly likely, neither likely nor unlikely, slightly unlikely, moderately unlikely, extremely unlikely) to receive a vaccine were grouped together in a second group."*

[17] Figure 3, p. 7: www.europeansocialsurvey.org/sites/default/files/2023-06/TL7-Immigration-English.pdf.

Yet a while ago, Alan Agresti, the renowned statistician most prolific on the analysis of ordinal categorical data, had already warned against such practice in the following terms: *"Some researchers collapse ordinal responses to binary so they can use ordinary logistic regression. However, a loss of efficiency occurs in collapsing ordinal scales, in the sense that larger standard errors result. In practice, when observations are spread fairly evenly among the categories, the efficiency loss is minor when you collapse a large number of categories to about four categories. However, it can be severe when you collapse to a binary response. It is usually inadvisable to do this."* (2007, p.185). Bultez et al. (2025) point out that such simplifying grouping of response categories persists in marketing research. In the section entitled: *Tacking stock of the calibration of ordinal categorical variables*, of their article**,** they sifted contributions archetypic of tests of relationships between consumers' satisfaction motives (product attributes, purchase experience), and their attitudes (preferences) and behavioral intentions (willingness to recommend, or to remain loyal) toward brands (cf. their Table 1). From this review, they conclude that: *"the complexity of cumulative multinomial regression has hampered its adoption."* (op. cit., § *Assumed cardinality*). For that reason, Alain Bultez programmed the tutorial: ***CATORDREG.xlxs***, designed to help those who want to better understand the econometrics behind the ***MCL*** model when predictors are also of an ordinal-categorical nature. This training tool is available online through ***ResearchGate***.

5. Empirical fitting of the *MCL* model to delivery time measurements

Hereafter, we essentially focus on the calibration of the $\boldsymbol{MCL}$ model that best fit the 2023 ***UNEX™-CEN*** data. Indeed, as evidenced by § 4.2.4., few publications document applications of the $\boldsymbol{MCL}$ regression: out of the numerous cases we reviewed, McCullagh (1980), Anderson and Philips (1981), Agresti (2007, pp. 182-184) stand out as notable exceptions … Not astonishing since these scholars pioneered and advocated the method. A few more can be found in epidemiology and medicine: e.g., Armstrong and Sloan (1988), Harrell et al. (1998), and in marketing: Bultez et al. (op. cit.). Anyway, the publication closest to our purpose we could find (not referred to here above because neither count, nor logistic regression was made used of by its authors) aimed at assessing 20 US commercial airports' operational efficiency through linear ordinary least-square regression analysis of flights' on-time arrival rates at these airports over a nine-year 2009-2017 period (Dinler and Rankin, 2020, Table 4, p.8). Hence, we must be among the first to test the ***MCL*** model on managerially relevant data and to show empirically how well it stands up against the ***NB*** model. In what follows, we argue why the ***MCL*** is as worth considering as count models for studying completion times and delays.

5.1. Estimation criterion

Let $\mathbb{S}$ be the large sample of delivery times $(\mathbb{S})$ measured in 2023, through the ***UNEX™*** experimentation system: $\mathbb{S} = \{t_1, t_2 \dots, t_i, \dots, t_n\}$, with: $n = |\mathbb{S}| = 105{,}889$. Such a data set can be readily processed through count regression, but to also apply the multinomial cumulative logistic regression, each record was encoded into a vector of $\mathbb{L} = (\bar{\mathbb{t}} + 1)$ binary dummies, when $\mathbb{L}$ stands for the category pooling late deliveries (i.e., all $t_i > \bar{\mathbb{t}}$):

$$t_i \Longrightarrow \left[\, y_{i,1}, y_{i,2}, \dots, y_{i,\mathbb{t}}, \dots, y_{i,\bar{\mathbb{t}}}, y_{i,\mathbb{L}}\right], \text{where: } y_{i,\mathbb{t}} = \begin{cases} 1, if\ t_i = \mathbb{t} \\ 0, if\ t_i \neq \mathbb{t} \end{cases}, y_{i,\mathbb{L}} = \begin{cases} 1, if\ t_i \geq \mathbb{L} \\ 0, if\ t_i < \mathbb{L} \end{cases} \text{and such that: } \sum_{\mathbb{t}=1}^{\mathbb{t}=\mathbb{L}} y_{\mathbb{t}} = 1.$$

The probabilistic nature of the models developed to infer indicators of delivery punctuality in section 4 (here above) justifies using the maximum likelihood method to parameterize them. Assuming observations are independent, conditional on the random effects (v^O, v^D), the conditional likelihood $\mathcal{L}(\mathbb{S}|v^O, v^D)$ of sample $\mathbb{S}$ is the product of the probabilities of the observed delivery times: $\pi_{i,\mathbb{t}}$, determined for fixed values of v_o^O and v_d^D.

Then, for the ***MCL*** model: $\pi_{t_i} \equiv \pi_{i,1}^{y_{i,1}} . \pi_{i,2}^{y_{i,2}} \ldots . \pi_{i,\bar{\mathbb{t}}}^{y_{i,\bar{\mathbb{t}}}} . \pi_{i,\mathbb{L}}^{y_{i,\mathbb{L}}}$, and

$$\mathcal{L}(\mathbb{S}|\boldsymbol{v^O}, \boldsymbol{v^D}) = \begin{cases} \prod_{i=1}^{i=n} \pi_{t_i} : \text{from compact count form } (\boldsymbol{NB} \text{ model }) & (\mathcal{L}.1) \\ \prod_{\boldsymbol{i=1}}^{\boldsymbol{i=n}} \prod_{\boldsymbol{t=1}}^{\boldsymbol{t=\mathbb{L}}} \pi_{i,t}^{y_{i,t}} : \text{from extended multinomial form } (\boldsymbol{MCL} \text{ model}) & (\mathcal{L}.2) \end{cases}$$

Table **5** details how this function is to be built. Advanced numerical techniques are required to integrate the conditional likelihood over random effects and maximize the resulting marginal likelihood. We opted for maximizing its Laplace approximation[18].

Table 5. Response records and their probabilities of occurrence according to the ***MCL*** model.

Performance classes	Delivery time: $\mathbb{t}$	Coding of records	Probability of delivery on day: $\mathbb{t}$
$\boldsymbol{\Theta_1 < \widetilde{Q}_i}$	1	$y_{i,1} = 1;\ y_{i,d} = 0, \forall d \neq 1$	$\pi_{i,1} = 1 - F(\theta_1 - q_i)$
$\boldsymbol{\Theta_2 < \widetilde{Q}_i \leq \Theta_1}$	2	$y_{i,2} = 1;\ y_{i,d} = 0, \forall d \neq 2$	$\pi_{i,2} = F(\theta_1 - q_i) - F(\theta_2 - q_i)$
...	...	...	...
$\boldsymbol{\Theta_{\mathbb{t}} < \widetilde{Q}_i \leq \Theta_{\mathbb{t}-1}}$	$\mathbb{t}$	$y_{i,\mathbb{t}} = 1;\ y_{i,d} = 0, \forall d \neq \mathbb{t}$	$\pi_{i,\mathbb{t}} = F(\theta_{\mathbb{t}-1} - q_i) - F(\theta_{\mathbb{t}} - q_i)$
...	...	...	...
$\boldsymbol{\Theta_{\bar{\mathbb{t}}} < \widetilde{Q}_i \leq \Theta_{\bar{\mathbb{t}}-1}}$	$\bar{\mathbb{t}}$	$y_{i,\bar{\mathbb{t}}} = 1;\ y_{i,d} = 0, \forall d \neq \bar{\mathbb{t}}$	$\pi_{i,\bar{\mathbb{t}}} = F(\theta_{\bar{\mathbb{t}}-1} - q_i) - F(\theta_{\bar{\mathbb{t}}} - q_i)$
$\boldsymbol{\widetilde{Q}_i \leq \Theta_{\bar{\mathbb{t}}}}$	$\mathbb{L}$	$y_{i,\mathbb{L}} = 1;\ y_{i,d} = 0, \forall d > \bar{\mathbb{t}}$	$\pi_{i,\mathbb{L}} = F(\theta_{\bar{\mathbb{t}}} - q_i)$

5.2. Capping the categorization

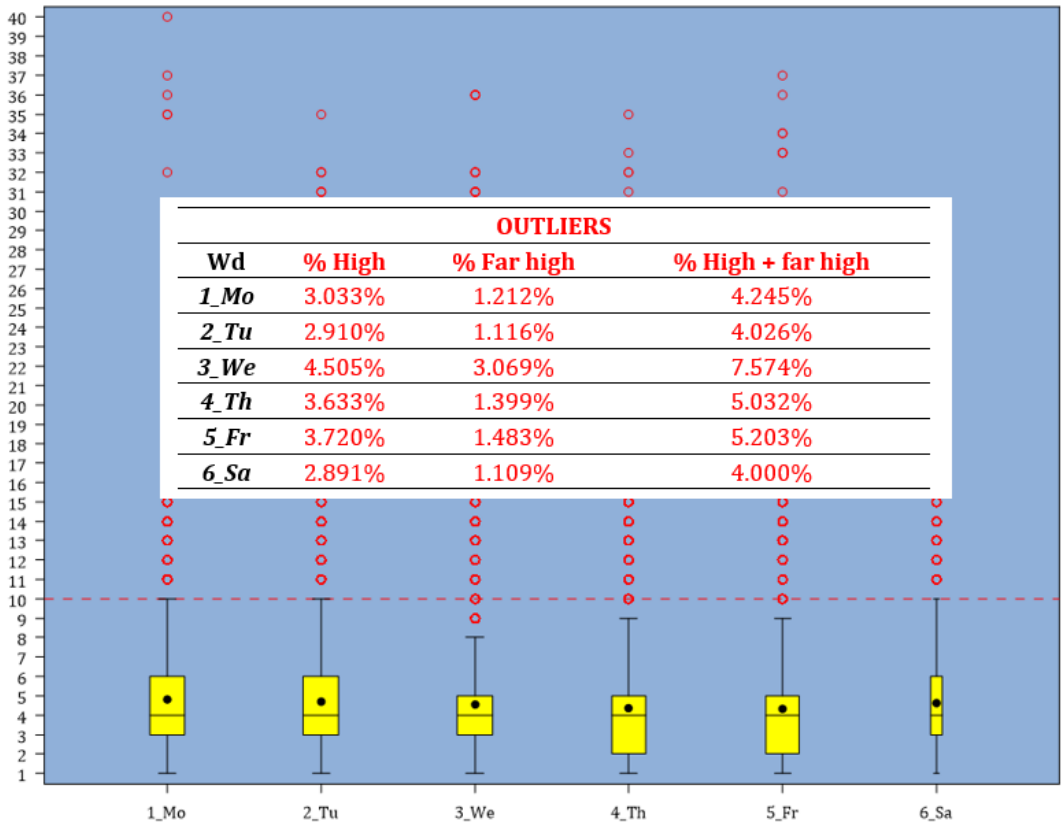
While the more parsimonious ***NB*** pmf extends over the semi-open interval: $\tilde{T} \in [0, \infty[$, the ***MCL*** cannot deal with an infinite range of logistic effectiveness grades because the $\boldsymbol{\Theta}_{\mathbb{t}}$-thresholds need to be parametrized. Hence, $\mathbb{L}$ must not be too large to avoid overfitting. Prior checking of the empirical distribution of measurements can aid in setting this cutoff.

Figure **5** - which draws in parallel the distributions of delivery times recorded by ***UNEX™*** in 2023, weekday per weekday - is quite helpful for that very purpose. Such a splitting is meant to neutralize variations caused by one the topmost discriminant fixed factors[19]: i.e., Wd. Each box-width is proportional to the square-root of the represented subsample size. Noticing that delivery times longer than 10 days pop up as abnormal delays[20], we naturally set $\mathbb{L}$ at 11, thus limiting the influence of outlying transit times represented by the red squares standing higher than the upper fence, thus located beyond the 1.5 the interquartile range above the 75th percentile.

[18] https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_glimmix_a0000001426.htm

[19] See: Tables **6** and **7**, hereafter.

[20] For mail collected on *Wednesday*, *Thursday* and *Friday*. the upper fence is located lower: at 8 (*We*) and 9 (*Th, Fr*).



OUTLIERS			
Wd	% High	% Far high	% High + far high
1_Mo	3.033%	1.212%	4.245%
2_Tu	2.910%	1.116%	4.026%
3_We	4.505%	3.069%	7.574%
4_Th	3.633%	1.399%	5.032%
5_Fr	3.720%	1.483%	5.203%
6_Sa	2.891%	1.109%	4.000%

Figure 5. Boxplots of delivery time records.

It comes as no surprise that within each rectangular yellow box the mean (black dot) stands above the median (horizontal segment) confirming the positive skewness of each distribution: right tail longer and fatter than the left.

5.3. Effects

Annex **A2** lists and comments the *SAS DATA STEP* and *LOGISTIC* procedure we adapted to parameterize the MCL model on sample $\mathbb{S}$, through the maximization of the approximation of the marginal distribution of the data derived by integration of $(\mathcal{L}.2)$ over the random components.

5.3.1. Across-the-board discriminative power of covariates

Tables **6** and **7** summarize the main outputs from fitting the multinomial cumulative logistic and negative binomial regression models to $\mathbb{S}$. Therein, the factors are ranked – from left to right – in decreasing order of statistical significance [21] of their global effect on delivery performance: reflecting the extent to which the parameters differentiating their modes depart from zero. More formally – by reference to the predictor function (5) –, each of the F-statistics jointly tests:

$$H_0: \bigcap_{m(f)\in\{\boldsymbol{M}^f | m(f)\neq b(f)\}} \left[\delta^f_{m(f)} = 0\right]$$

$$\text{against } H_1: \bigcup_{m(f)\in\{\boldsymbol{M}^f | m(f)\neq b(f)\}} \left[\delta^f_{m(f)} \neq 0\right], \qquad for\ f \in \boldsymbol{\mathcal{F}}.$$

[21] That is, by increasing order of associated p-values.

Table 6. MCL regression output.

FACTOR: f	Co	Co	Wd	Fk	Pl	Uo	Sw	Ud
F**-statistic (type:** III**)**	94.09	86.95	476.12	13.44	11.00	3.27	2.16	1.85
$v_{Numerator}$	30	30	5	2	2	2	2	2
p**-value**	0.00%	0.00%	0.00%	0.00%	0.00%	3.82%	11.48%	15.71%

Variances of random components: v_o^O and v_d^D, reflecting local random variations in logistics

Ends	**Outbound:** $[\varsigma^O]^2$	**Inbound:** $[\varsigma^D]^2$
Estimates	0.0294	0.0327
Standard errors	0.0051	0.0059
Precision: ratio of the estimate to its standard error	5.76	5.57

Table 7. NB regression output.

FACTOR: f	Co	Co	Wd	Fk	Pl	Sw	Uo	Ud
F**-statistic (type:** III**)**	92.12	63.41	185.03	13.49	4.52	2.27	1.7	0.90
$v_{Numerator}$	30	30	5	2	2	2	2	2
p**-value**	0.00%	0.00%	0.00%	0.00%	1.09%	10.35%	18.32%	40.83%

Variances of random components: v_o^O and v_d^D, reflecting local random variations in logistics

Ends	**Outbound:** $[\varsigma^O]^2$	**Inbound:** $[\varsigma^D]^2$
Estimates	0.0026	0.0037
Standard errors	0.0005	0.0006
Precision: ratio of the estimate to its standard error	5.51	5.80

OVERDISPERSION	***Estimate***	***Standard error***	***Precision***
Scale	$\hat{\phi} = 0.0976$	$SE[\hat{\phi}] = 0.0014$	$\hat{\phi}/SE[\hat{\phi}] = 70.10$

The relatively low significance of effect estimates of

- envelope size/weight (*Sw*) is due to advances in sorting technology and wider use of standardized trays and containers, which are better suited at handling both large and small formats,
- the degrees of urbanization (*Uo* and *Ud*) may be caused by the inclusion of the random components but also partly by unorthodox classifications: several posts use logics like geography or administrative zonal split of the territory.

At the light of the p-values, it appears that out of the eight fixed predictors, the effects of Co, Cd, Wd, Fk and Pl dominate those of the others, while Uo matters less, and Sw as well as Ud would be considered insignificant by those who blindly apply the below-5%-rule (severely criticized by, among others, Bultez et al., 2022). However, we kept them all in the model because:

a. A priori, they look relevant to Posts according to their expertise, therefore, they all determine the stratification design (refer to § 3.1.1) and with Agresti (2002, p. 214) we believe that *"it is sensible to include a variable that is central to the purposes of the study and report its estimated effect even if it is not statistically significant. Keeping it in the model may help reduce bias in estimated effects of other predictors".*

b. Following suit hundreds of colleagues from all disciplines (Amrhein et al., 2019), we have banned the mechanical dichotomization of effects' estimates into *significant* versus *nonsignificant* ones. We should be more sagacious and acknowledge that (Mulder and Wagenmakers, 2016, p.1): *"the p value does not allow one to discriminate absence of evidence (i.e., uninformative data) from evidence of absence (i.e., data supporting the null hypothesis)"*.

c. *"Expressions of uncertainty"* such as p-values *"are themselves uncertain"* (Calin-Jageman and Cumming, 2019, p.277), conditioned, as they are, by the *"model specification, sample selection and the handling of data issues"*.

d. In line with point *c*, we tested the sensitivity of the empirical statistical criteria to the model specification by benchmarking the $\boldsymbol{MCL}$ against the negative binomial regression – i.e., (10)-(12) and integral of $(\mathcal{L}.1)$. The *SAS GLIMMIX* procedure used to perform this additional analysis is listed and annotated in annex **A.3**. Results displayed in Table **7** reveal markedly lower F-statistic values for Co, Wd, Pl, Uo and Ud. This proves the lack of robustness of diagnoses relying solely on such indicators and related p-values, which depend on the model (*c*).

5.3.2. Modal differentiation

On top of the four reasons (a to d) why we want to keep the specification as complete as possible and are reluctant to rely on the sole p-values to sort out allegedly negligible covariates, one should keep in mind that *"larger p-values do not imply a lack of importance or even lack of effect … large effects may produce unimpressive p-values if the … measurements are imprecise"* (Wasserstein and Lazar, 2016, p. 132). Therefore, specific mode-level *effect-sizes* deserve special attention: Bultez and Herrmann illustrated their relevance through their review of recent publications in marketing (2025, sections: *The null hypothesis matters* and *From significance to salience*).

To gauge items' attributes influence on their handling by the posts, let us consider, as in (20), two of them: $i \neq \ell$, but solely differentiated by one factor, say f: i.e., $m(f)$ for $i \neq \mathfrak{m}(f)$ for ℓ. So,

$$x^f_{i,m(f)} = x^f_{\ell,\mathfrak{m}(f)} = 1, \text{while: } x^{\not f}_{i,m(\not f)} = x^{\not f}_{\ell,m(\not f)}, \forall m(\not f) \in \boldsymbol{M^{\not f}}, \forall \not f \neq f. \tag{21.1}$$

Then, according to (15.3), this unique dissimilitude results in the following *QoS* differential:

$$q_i - q_\ell = \sum_{m(f)\in\{M^f | m(f)\neq b(f)\}} \delta^f_{m(f)/\sigma}.\left(x^f_{i,m(f)} - x^f_{\ell,m(f)}\right) = \delta^f_{m(f)/\sigma} - \delta^f_{\mathfrak{m}(f)/\sigma},$$

or, using (5.1),

$$q_i - q_\ell = \left[\left(\beta^f_{m(f)} - \beta^f_{b(f)}\right) - \left(\beta^f_{\mathfrak{m}(f)} - \beta^f_{b(f)}\right)\right]/\sigma = \left(\beta^f_{m(f)} - \beta^f_{\mathfrak{m}(f)}\right)/\sigma.$$

Thus, mode $m(f)$ can be contrasted with mode $\mathfrak{m}(f)$ by assessing to what extent

- either the standardized difference between related parameters departs from zero:

$$\Delta^f_{m(f),\mathfrak{m}(f)} = \left(\beta^f_{m(f)} - \beta^f_{\mathfrak{m}(f)}\right)/\sigma = 0 \Longrightarrow q_i = q_\ell, \tag{21.2}$$

- or the corresponding *odds ratio* deviates from 1, since (20) implies:

$$q_i = q_\ell \Longrightarrow OR_{i,\ell} \equiv OR^f_{m(f),\mathfrak{m}(f)} = e^0 = 1. \tag{21.3}$$

Henceforth, if $\Delta^f_{m(f),\mathfrak{m}(f)} \cong 0$, or $OR^f_{m(f),\mathfrak{m}(f)} \cong 1$, modes $m(f)$ and $\mathfrak{m}(f)$ need not be singled out.

Both (21.2) and (21.3) materialize the ceteris paribus *QoS* differentials in the delivery of letters characterized by dissimilar modes of a single predictor. They inform about the possibilities of pooling these modes … gradually and prudently because, as Goodman et al. (2019, p. 170) rightly pointed it, *"there is no fixed answer for how large a difference must be from the null to be considered meaningful"*. In fact, the substantiveness of effects can only be assessed contextually, against a *"meaningful"* benchmark, but *"in practice"* fixing such a reference value *"is complicated*", and having it endorsed by stakeholders who don't share common views is even trickier (Betensky, 2019, pp. 115-116). Therefore, we prefer for ***MCL*** models appreciating mode-level effect-sizes in relative probabilistic terms, through odds ratios, contrasted with those observed for other relevant modes. For that purpose, plots parallelizing their confidence intervals facilitate the comparisons of the magnitudes and imprecisions of effects' measurements. Such forest charts were popularized in medicine to sum up reviews of experimental results: e.g., Harrell (2015, pp. 281-282) alternates between log-odds (op. cit., Figure 11.2) and *"intervals drawn on the log odds ratio scale but labelled on the odds ratio scale"* (op. cit., Figure 11.3). Figure **6** illustrates how such forest charts facilitate the assessment of the relative magnitudes of class effects underlying a nominal predictor. Here, it highlights weekdays' differentials (Wd), i.e. the most discriminant factor in 2023, as shown by Tables **6** and **7.** Its ordinates (vertical axis) are meaningless: they simply correspond to the ordering of the legend).

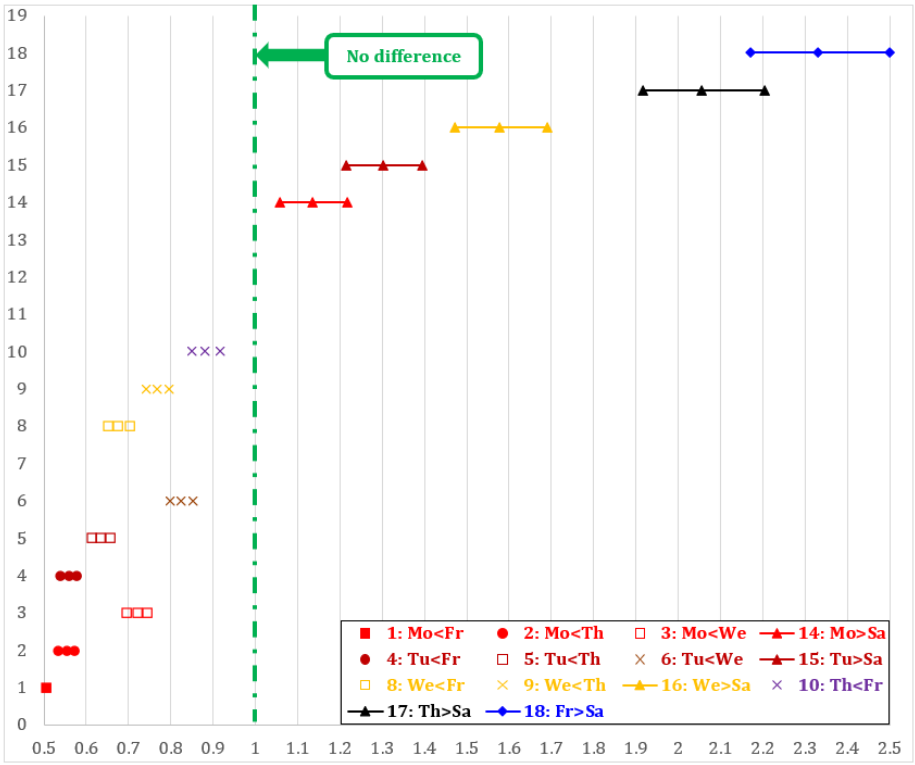


Figure 6. 95% Confidence intervals of odds ratios contrasting weekdays on which letters get posted: $OR^{Wd}_{m(Wd),\mathfrak{m}(Wd)}$, with $m(Wd) \neq \mathfrak{m}(Wd) \in \{Mo, Tu, We, Th, Fr, Sa\}$.

Interval positioning and extents are to be judged along the horizontal axis[22].

[22] One should be cautious when comparing the magnitudes of odds ratios inferred from logistic regression on different samples, or from different models, for (21.2) reminds us that modes' differentials are scaled in σ-unit: *"Different odds ratios from the same study cannot be compared when the statistical models that result in odds ratio estimates have different explanatory variables because each model has a different arbitrary scaling factor. Nor can the magnitude of the odds ratio from one study be compared with the magnitude of the odds ratio from another study, because different samples and different model specifications will have different arbitrary scaling factors. A further implication is that the magnitudes of odds ratios of a given association in multiple studies cannot be synthesized in a meta-analysis"* (Norton et al., 2018, p. 84).

Figure **6** makes it clear that grouping mailing days is out of question: none of the segments of estimates most compatible with $\mathbb{S}$ is close to overlapping the no-difference line. The nearest one is $\widehat{OR}^{Wd}_{Mo,Sa}$ and it implies: $\hat{\beta}^{Wd}_{Mo} > \hat{\beta}^{Wd}_{Sa}$. All those located to the right of the demarcation involve $Saturday$ and are such that:

$$\widehat{OR}^{Wd}_{Mo,Sa} < \widehat{OR}^{Wd}_{Tu,Sa} < \widehat{OR}^{Wd}_{We,Sa} < \widehat{OR}^{Wd}_{Th,Sa} < \widehat{OR}^{Wd}_{Fr,Sa},$$

- letters mailed on $Saturday$ are handled less efficiently than those posted on any other day:

$$\beta^{Wd}_{m(Wd)} > \beta^{Wd}_{Sa}, \forall m(Wd) \neq Sa,$$

- the ranking of weekdays' coefficients in increasing order logistic efficacy corresponds to the chronology:

$$\beta^{Wd}_{Mo} < \beta^{Wd}_{Tu} < \beta^{Wd}_{We} < \beta^{Wd}_{Th} < \beta^{Wd}_{Fr}.$$

The widest confidence intervals also materialize contrasts with $Saturday$, indicating that delivery time of letters posted on weekends is highly uncertain.

5.4. QoS thresholds

Intercepts in the ***MCL*** model define critical performance levels along the *QoS* continuum. Thereby, these unknown constants contribute to determining the probabilities of delivery within deadlines. Thus, from analyzing gaps between them, much can be learned about efficiency efforts to be put in delivery operations. Therefore, it is worth pondering to what extent the values of these cut-points grow as the cutoff dates for receiving letters get closer to their posting day.

5.4.1. Uneven spurts

Successive jumps in thresholds can be statistically assessed by testing the following hypotheses:

$$H_0: \Theta_{\mathbb{t}} = \Theta_{\mathbb{t}+1} \text{ versus } H_1: \Theta_{\mathbb{t}} > \Theta_{\mathbb{t}+1}, \text{for } \mathbb{t} \in \{1,2,3,4,5,6,7,8,9\},$$

or equivalently, $H_0: \theta_{\mathbb{t}|B} - \theta_{\mathbb{t}+1|B} = 0$ versus : $\theta_{\mathbb{t}|B} - \theta_{\mathbb{t}+1|B} > 0$, for $\mathbb{t} \in \{1,2,3,4,5,6,7,8,9\}$,

because according to (15.4), $\theta_{\mathbb{t}|B} = \theta_{\mathbb{t}} - (\mathbf{B}/\sigma) \Longrightarrow \theta_{\mathbb{t}|B} - \theta_{\mathbb{t}+1|B} = \theta_{\mathbb{t}} - \theta_{\mathbb{t}+1} = (\Theta_{\mathbb{t}}/\sigma) - (\Theta_{\mathbb{t}+1}/\sigma)$.

Under *SAS*, these tests can be programmed using the *CONTRAST*-statements of the *GLIMMIX* procedure, specified as in paragraph ***II.2*** of annex **A2**. The outputs from the execution of these statements are, however, too succinct: only the z^2- and p-values assessing the significance of the discrepancies are produced by *SAS*. Hence, we programmed those comparative tests through the %***Cut_points***-MACRO listed in section **III** of annex **A2**. Table **8** reports the output from the execution of this MACRO: as p-values are all extremely close to zero (for $z > 5$, the one-tail p is lower than $2.87\text{x}10^{-7}$), they are not displayed. Thus, as all differences are highly significantly positive: H_1 holds true. Focusing on the third column, reading the $\widehat{\boldsymbol{\Delta}}_{\mathbb{t},\mathbb{t}+1}$-values bottom-up, one realizes that ***the higher the performance level reached (smaller*** $\mathbb{t}$***)***, the larger the difference between estimated thresholds (up the ladder) becomes, ***the harder it is to shave a day off the delivery time.*** In other words, the higher the *QoS*-threshold is, the lower the probability of speeding up the process by one day:

$$\theta_{\mathbb{t}} \gg \theta_{\mathbb{t}+1} \Longrightarrow P(\tilde{Q}_i > \Theta_{\mathbb{t}}) \ll P(\tilde{Q}_i > \Theta_{\mathbb{t}+1}) \Longrightarrow P(\tilde{\mathbb{T}}_i \leq \mathbb{t}) \ll P(\tilde{\mathbb{T}}_i \leq \mathbb{t}+1).$$

5.4.2. Double jeopardy

The consequences of ever-rising $\widehat{\boldsymbol{\Delta}}_{\mathbb{t},\mathbb{t}+1}$-steps exhibited in Table **8** can be more precisely inferred from (19):

$$ln(\mathbb{O}(\mathbb{T}_i \leq \mathbb{t})/\mathbb{O}(\mathbb{T}_i \leq \mathbb{t}+1)) = (q_i - \theta_{\mathbb{t}|\mathrm{B}}) - (q_i - \theta_{\mathbb{t}+1|\mathrm{B}}) = -(\theta_{\mathbb{t}|\mathrm{B}} - \theta_{\mathbb{t}+1|\mathrm{B}}) = -(\theta_{\mathbb{t}} - \theta_{\mathbb{t}+1}).$$

This formula establishes that *the ratio of odds in favour of the shorter delivery time declines exponentially with the QoS ladder rung spacing*:

$$\boldsymbol{OR}^{\mathbb{t}}_{\mathbb{t}+1} = \mathbb{O}(\mathbb{T}_i \leq \mathbb{t})/\mathbb{O}(\mathbb{T}_i \leq \mathbb{t}+1) = e^{-(\theta_{\mathbb{t}} - \theta_{\mathbb{t}+1})} = e^{-\Delta_{\mathbb{t},\mathbb{t}+1}}. \quad (22)$$

The last column of Table **8** and Figure **7** empirically demonstrate the conjunction of escalating $\widehat{\boldsymbol{\Delta}}_{\mathbb{t},\mathbb{t}+1}$-jumps with the non-linearity in the fall of favourable odds.

Table 8. Increasing jumps in *QoS* required to enhance probability of delivery within a shortened deadline.

t	$\widehat{\theta}_{t\|B}$	$\widehat{\Delta}_{t,t+1}$: $\widehat{\theta}_t - \widehat{\theta}_{t+1}$	Standard error of $\widehat{\Delta}_{t,t+1}$: $SE[\widehat{\Delta}_{t,t+1}]$	Standardized difference: z-statistic $= \widehat{\Delta}_{t,t+1}/SE[\widehat{\Delta}_{t,t+1}]$	$\widehat{OR}^{t}_{t+1} = e^{-\widehat{\Delta}_{t,t+1}}$
1	10.532				
2	7.368	**3.164**	0.0246	128.40	**0.042**
3	5.861	**1.507**	0.0089	169.72	**0.221**
4	4.779	**1.083**	0.0072	150.73	**0.339**
5	3.913	**0.865**	0.0069	125.76	**0.421**
6	3.204	**0.709**	0.0071	100.04	**0.492**
7	2.622	**0.582**	0.0075	77.89	**0.559**
8	2.142	**0.481**	0.0079	60.70	**0.618**
9	1.717	**0.425**	0.0086	49.15	**0.654**
10	1.373	**0.344**	0.0089	38.50	**0.709**

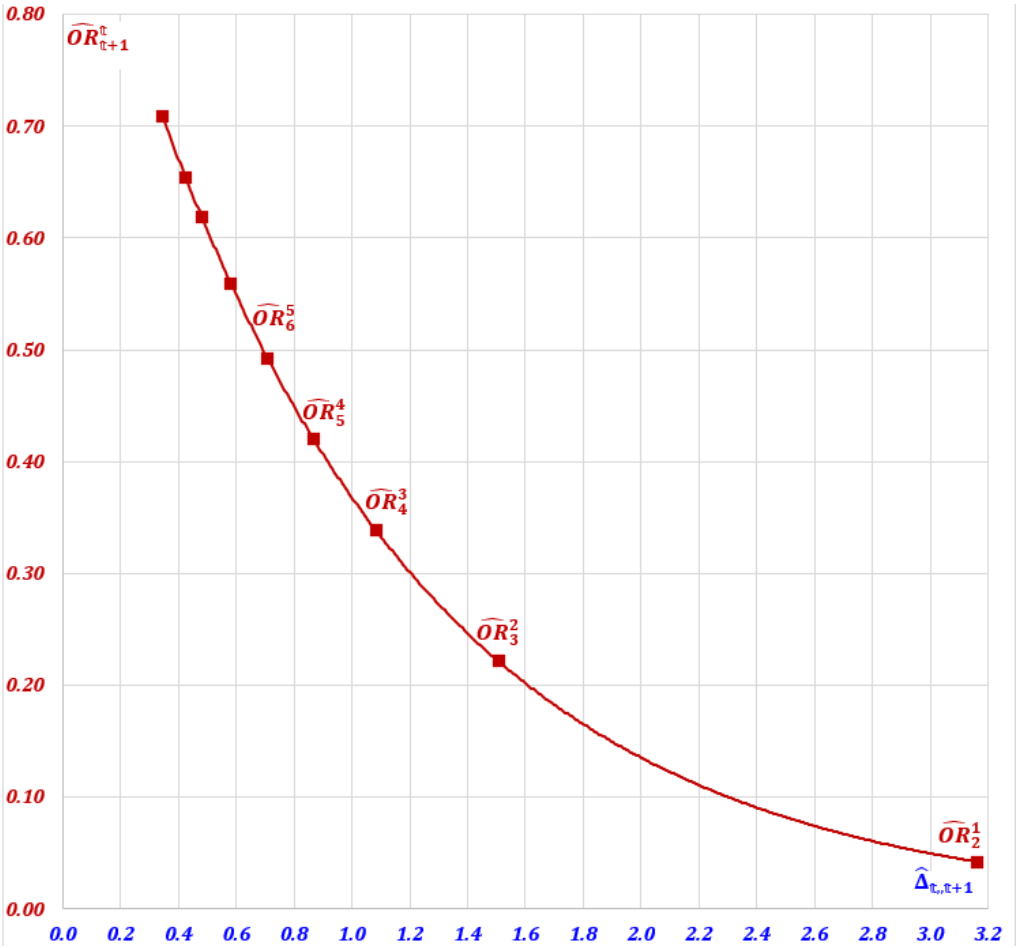


Figure 7. Drops in the odds for a one-day faster delivery time resulting from increasing leaps in *QoS* thresholds.

5.4.3. Lead-time shrinking comes at a price

The efficiency gain needed to win a day can best be realized by visualizing the failure-to-achieve-target ***risks,*** which according to (18.2) are to be estimated by:

$$\hat{R}(\mathbb{t}|\hat{q}) = \hat{P}(\tilde{\mathbb{T}} > \mathbb{t}|\hat{q}) = F(\hat{\theta}_{\mathbb{t}} - \hat{q}). \tag{23}$$

Note that, while subscript i is now superfluous - since for predictions of response functions there is no need to distinguish items -, one must make clear that such probability of flop is conditioned by the *QoS*-level. Figure **8** maps the risk-curves for $\mathbb{t} \in \{1,2,3,4,5,6,7,8,9,10\}$ generated by (23), using the $\hat{\theta}_{\mathbb{t}}$ values reported in Table **7**, one curve per target $\mathbb{t}$.

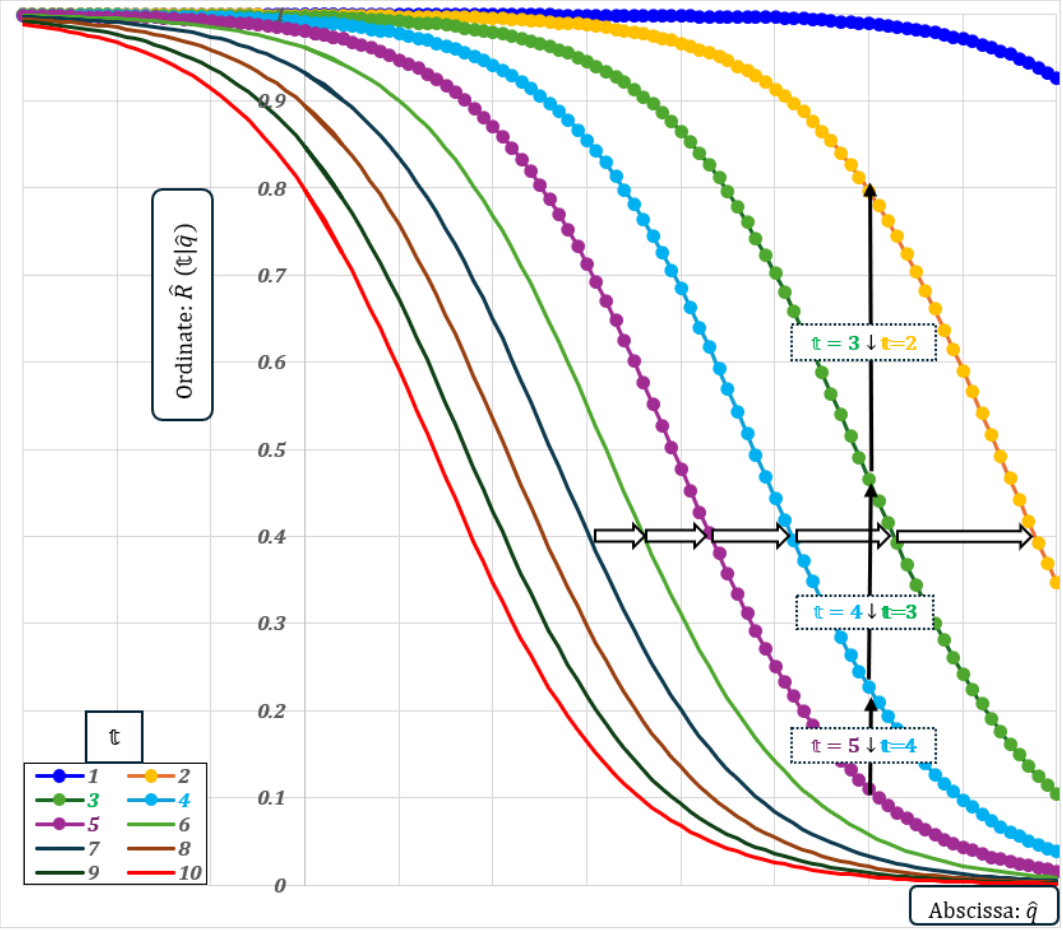


Figure 8. Risks of missing deadlines decrease as *QoS* improves.

This graph illustrates that for every deadline, superior *QoS* levels (plotted along the horizontal axis) entail lower risks of delivery beyond the targeted deadline (plotted along the vertical axis); tighter deadlines translate into higher risks: the smaller $\mathbb{t}$, the upper the curve. Thus, the difficulty of delivering within shorter time spans increases exponentially. Eyeballing this graph, horizontally from left to right, one sees that to ensure a fixed degree of risk, efficiency gains to be achieved to reach more stringent targets escalate at an accelerating rate: e.g., the sizes of non-filled black-block arrows at the ordinate $\hat{R}(\mathbb{t}|\hat{q}) = 0.4$ patently stretch as $\mathbb{t}$ shrinks. The dual reality is as obvious, since examining enlarging gaps, bottom-up along the black arrows, one realizes that jumps in the risk of failure become greater and greater:

$$[\hat{R}(9|\hat{q}) - \hat{R}(10|\hat{q})] < \cdots [\hat{R}(5|\hat{q}) - \hat{R}(6|\hat{q})] < [\hat{R}(4|\hat{q}) - \hat{R}(5|\hat{q})] < [\hat{R}(3|\hat{q}) - \hat{R}(4|\hat{q})] < [\hat{R}(2|\hat{q}) - \hat{R}(3|\hat{q})].$$

*5.5. Confirmation by the **NB**, from two other outlooks*

Thus far, from the ***MCL*** model optimal fitting, we inferred risks of missing due dates: $\hat{R}(\mathbb{t}|\hat{q}, \boldsymbol{MCL})$. Still, the probability mass function defining chances of delivery of a letter exactly $\mathbb{t}$ days after its posting may teach us even more. According to formula (16.1), these chances result from the following differences:

$$\hat{\pi}_{\mathbb{t}|\hat{q},MCL} = \hat{P}(\tilde{\mathbb{T}} = \mathbb{t}|\hat{q}, \boldsymbol{MCL}) = \Lambda(\hat{q} - \hat{\theta}_{\mathbb{t}}) - \Lambda(\hat{q} - \hat{\theta}_{\mathbb{t}-1}). \quad (24)$$

Rather, should one calibrate the ***NB*** model, equations (10) and (11) would apply to yield another set of estimates:

$$\hat{\pi}_{\mathbb{t}|\hat{Q},NB} = \hat{P}(\tilde{\mathbb{T}} = \mathbb{t}|\hat{Q}, \boldsymbol{NB}) = \frac{\Gamma(\mathbb{t} - 1 + \hat{\phi}^{-1})}{\Gamma(\hat{\phi}^{-1}).\Gamma(\mathbb{t})} . \hat{p}^{(1/\hat{\phi})}.(1 - \hat{p})^{\mathbb{t}-1}, \text{with: } \hat{p} = \frac{1}{1 + (\hat{\phi}.\hat{\mu})} \text{ and } \hat{\mu} = e^{-\hat{Q}}. \quad (25)$$

In (24) and (25), the model label is added in the condition specification, to make clear that we don't expect that: $\hat{\pi}_{\mathbb{t}|\hat{Q},NB} = \hat{\pi}_{\mathbb{t}|\hat{q},MCL}$. Figure **9** synthesizes the distributions of probabilities estimated by (24) and (25), for $\mathbb{t} = 1$ up to $\mathbb{t} = 5$, using the same colours as in Figure **8**.

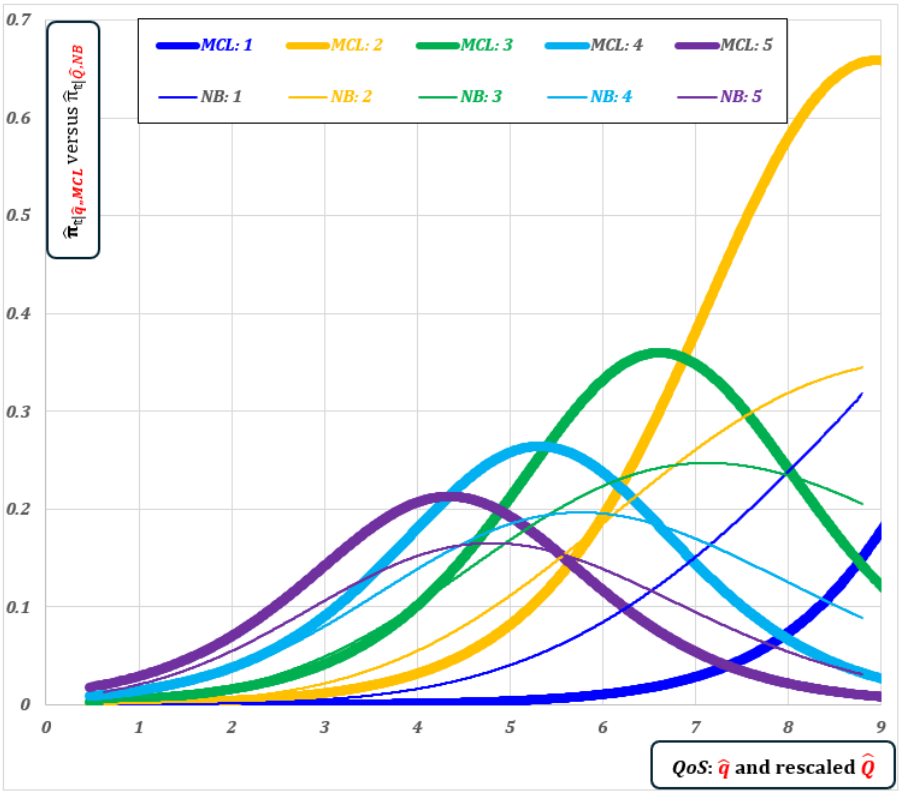


Figure 9. Polytomous distributions of predicted probabilities of delivery $\mathbb{t}$ days after posting.

On this chart, the $\hat{\pi}_{\mathbb{t}|\hat{q},MCL}$- and $\hat{\pi}_{\mathbb{t}|\hat{Q},NB}$-curves are contrasted by lines with different widths: the thicker ones for the $\hat{\pi}_{\mathbb{t}|\hat{q},MCL}$. They result from plotting models' fitted values for the 51,869 effectively surveyed strata, according to identically scaled ordinates, while their predictors - i.e., the estimates of the latent *QoS* - are to be read in abscissa. Note that as the two models generate dissimilar *QoS* scores: $\hat{Q} \neq \hat{q}$, we rescaled $\hat{Q}$ by a linear transformation (determined by regressing $\hat{Q}$ on $\hat{q}$)[23] to warrant comparability of the $\hat{\pi}_{\mathbb{t}|\hat{Q},NB}$ with the $\hat{\pi}_{\mathbb{t}|\hat{q},MCL}$. Both sets indicate how chances of faster shipping escalate at the expense of probabilities of slower dispatching as *QoS* gets enhanced. Except for the shortest delivery time (i.e., $\mathbb{t} = \mathbf{1}$), all curves increase first and, after culminating

[23] $\hat{Q}$ is relatively highly positively correlated with $\hat{q}$: $R = 0.96$ (regression coefficient: 3.36, intercept: 9.45). Although this rescaling of $\hat{Q}$ into $\hat{q}$-units is imperfect, it enables us to plot on the same graph (Figures **9** and **10**) the response probabilities inferred from the two models.

at a peak (beyond the limits of the graph for $\mathbb{t} = \mathbf{2}$), decrease. Over the entire range of estimated *QoS*-levels, $\hat{\pi}_{1|\hat{q},MCL}$ and $\hat{\pi}_{1|\hat{Q},NB}$ rise very slowly first and then rapidly while $\hat{\pi}_{\mathbb{L}|\hat{q},MCL}$ and $\hat{\pi}_{\mathbb{L}|\hat{Q},NB}$ (not displayed to make the graph more readable) fall all the way through quite sharply. Scanning the graphs from right to left, one notices that peaks for shorter delivery times (dominating on the right side) stand above, and are located to the right of, those for longer ones (dominated on the left), just because the better (worse) the *QoS*, the more likely shorter (longer) delivery times are. Shapes of the $\hat{\pi}_{\mathbb{t}|\hat{q},MCL}$ and $\hat{\pi}_{\mathbb{t}|\hat{Q},NB}$ curves are pretty much alike in their overall appearance. A closer look at them reveals that as the deadline gets shortened, peaks predicted by ***NB*** tend to detach more and more from those predicted by the ***MCL***. Table **9** strengthens this impression.

Table 9. Distances between peaks in probabilities of delivery $\mathbb{t}$ days after posting.

	Abscissas at which maxima get reached			
$\mathbb{t}$	$\hat{Q}^*(\mathbb{t},NB) = \underset{\hat{Q}}{\text{argmax}}\{\hat{\pi}_{\mathbb{t}\|\hat{Q},NB}\}$	$\hat{q}^*(\mathbb{t},MCL) = \underset{\hat{q}}{\text{argmax}}\{\hat{\pi}_{\mathbb{t}\|\hat{q},MCL}\}$	$\hat{Q}^*(\mathbb{t},NB) - \hat{Q}^*(\mathbb{t}+1,NB)$	$\hat{q}^*(\mathbb{t},MCL) - \hat{q}^*(\mathbb{t}+1,MCL)$
2		8.953		
3	7.127	6.615		2.338
4	5.765	5.320	1.361	1.295
5	4.799	4.346	0.966	0.974

	Peak ordinates			
$\mathbb{t}$	$\hat{\pi}^*_{\mathbb{t}\|\hat{Q}^*,NB}$	$\hat{\pi}^*_{\mathbb{t}\|\hat{q}^*,MCL}$	$\hat{\pi}^*_{\mathbb{t}\|\hat{Q}^*,NB} - \hat{\pi}^*_{\mathbb{t}+1\|\hat{Q}^*,NB}$	$\hat{\pi}^*_{\mathbb{t}\|\hat{q}^*,MCL} - \hat{\pi}^*_{\mathbb{t}+1\|\hat{q}^*,MCL}$
2		0.659		
3	0.247	0.360		0.299
4	0.197	0.264	0.051	0.096
5	0.165	0.213	0.031	0.051

More importantly, the ***NB*** model, like the ***MCL***, also captures the fact that reducing delivery time requires increasingly demanding endeavors in *QoS*, objectified by Figure **10**.

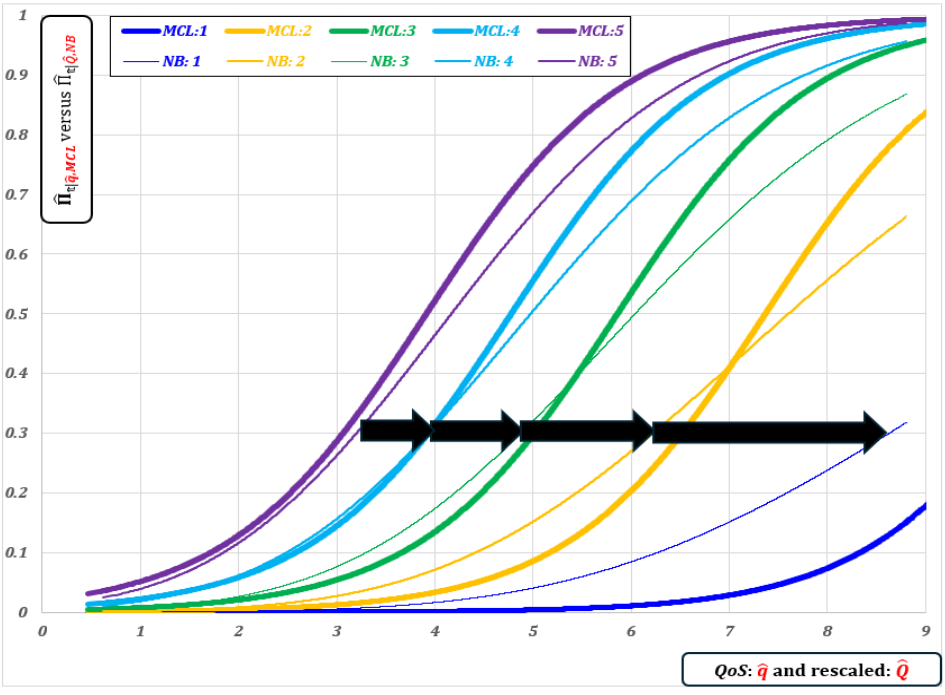


Figure 10. Enlarging gaps between *QoS*-levels to preserve chances of delivering by the due date.

Cumulative variant of Figure **9**, Figure **10** is the positive dual of Figure **8** since it gathers ***chances of punctual conveyance***, inferred from both models: $\widehat{\Pi}_{\mathbb{t}|\hat{q},MCL}$ and $\widehat{\Pi}_{\mathbb{t}|\hat{Q},NB}$, for the strata surveyed in 2023. It shows, once more, that as we go down from $\mathbb{t}=5$ to $\mathbb{t}=1$, one day at a time, the objective becomes more and more challenging, a point already made in § 5.4.2 and 5.4.3, but just for the ***MCL*** and from the ***risks of delays'*** angle. Distances between curves – stressed by the horizontal left-to-right filled block-black arrows separating the ***NB***-based estimates of ***chances to meet deadlines*** - measure the additional productivity needed to cut the delivery time by one more day, while maintaining the probability of intime delivery (here, at 0.3): the necessitated extra effort obviously raises with the level of effectiveness already attained. Thus, both models signal that some form of law of diminishing returns to investments in *QoS* applies to postal logistics: technological limitations are at work.

*5.6. Benchmarking **NB** against **MCL***

From a managerial viewpoint, the two models had better be compared by reviewing more closely predictions of stratum-level punctuality rates: i.e., the ***probabilities of delivery within fixed deadlines*** (i.e., by $\mathbb{t}$, at the latest) defined in section **4**. Figure **10** can only roughly[24] help visualize (dis)similarities between the those generated by the ***NB*** model against those produced by the ***MCL*** model; eyeballing it, we can just assert: *(a)* the ***NB*** overrates the ***MCL***-estimates of chances of delivery within one day, and so much so that *QoS* is high; *(b)* over-(under-)grading of chances of delivery within two days by the ***NB*** occurs at lower(upper) *QoS*-levels. Therefore, to get a more accurate picture we benchmarked one model's predictions against those provided by the second through linear regression equations, i.e. one equation per $\mathbb{t}$-deadline:

$$\widehat{\Pi}_{s,\mathbb{t}|MCL} = \kappa_0 + \kappa_1 . \widehat{\Pi}_{s,\mathbb{t}|NB} \text{ on the strata covered by } UNEX \text{ in 2023: } s \in \{1, 2, 3, \dots, nS = 51{,}869\}. \quad (26)$$

Table **10** sums up the results of these fits. It sheds light on the correlations between the two sets of predictions, as well as on the over(under)estimation of the regressand (***MCL***) by the regressor (***NB***).

Table 10. Relationships between model-predicted probabilities of delivery within deadlines.

Deadline: $\mathbb{t}$	Statistics	κ_0	κ_1	R^2	% Cases: $\widehat{\Pi}_{s,\mathbb{t}\|NB} \gtreqless \widehat{\Pi}_{s,\mathbb{t}\|MCL}$
1	Estimate (E)	-0.00547	0.26028	0.74598	Overestimation: *NB>>>MCL* 100.00%
	Standard error (SE)	0.00006	0.00067		
	T-ratio = E/SE	-95.01	390.28		
2	Estimate (E)	-0.05562	1.10323	0.86651	Overestimation: *NB>>MCL* 82.36%
	Standard error (SE)	0.00046	0.00190		
	T-ratio = E/SE	-119.87	580.23		
3	Estimate (E)	-0.06145	1.20669	0.91308	No trend: *NB≅MCL* 50.75%
	Standard error (SE)	0.00069	0.00163		
	T-ratio = E/SE	-89.76	738.17		
4	Estimate (E)	-0.01816	1.11758	0.92032	Underestimation: *NB<MCL* 68.98%
	Standard error (SE)	0.00083	0.00144		
	T-ratio = E/SE	-21.91	773.98		
5	Estimate (E)	0.04760	1.00551	0.91606	Underestimation: *NB<MCL* 81.11%
	Standard error (SE)	0.00093	0.00134		
	T-ratio = E/SE	51.09	752.33		

[24] Because of the non-exact match of the *QoS*-scores (cf. footnote 23 about the scale of the horizontal axis).

The coefficients of determination ($\boldsymbol{R}^2$) are reasonably high and estimates of $\boldsymbol{\kappa_1}$ are positive: as it should, correlations are positive and relatively strong. For $\mathbb{t} = \mathbf{1}$, $\boldsymbol{\kappa_0}$ is negative and $\boldsymbol{\kappa_1}$ is much smaller than 1 which means that for all strata, $\widehat{\Pi}_{s,1|\boldsymbol{NB}}$ largely overestimates $\widehat{\Pi}_{s,1|\boldsymbol{MCL}}$, which Figure **9** revealed: $\widehat{\pi}_{s,1|\boldsymbol{NB}} > \widehat{\pi}_{s,1|\boldsymbol{MCL}}$, and Figure **10** confirmed: $\widehat{\Pi}_{s,1|\boldsymbol{NB}} > \widehat{\Pi}_{s,1|\boldsymbol{MCL}}$ (blue curves: the first in the legend). From $\mathbb{t} = \mathbf{1}$ up to $\mathbb{t} = \mathbf{5}$, the overestimation (resp. underestimation) rate decreases (increases). For sure, the predictions from the ***NB*** model fall short of approximating those derived from the ***MCL***.

5.7. Takeaways

Winkelmann (2008, p. 68) rightly pointed out: *"…ordinal models can also be used for counts as long as the number of different counts observed in the sample is not too large. The number of threshold parameters that require estimation increases with the observed sample space by one-to-one … **Ordered models in general provide a better fit to the data than pure count data models. The threshold parameters give the flexibility to align predicted and actual frequencies.**"* He further noted that: "*However, their use for modeling count data has a number of serious deficiencies.*

- *They are theoretically implausible as a model for counts. They are not based on the concept of an underlying count process.*
- *Counts are cardinal rather than ordinal. Hence, under the ordinal approach, the sequence "2, 5, 50" is assumed to carry the same information as the sequence "0, 1, 2" which is not the case for count data. Ordinal models disregard this information and cannot be efficient.*
- *One reason of having parametric models in the first place is the ability of predicting the probability of arbitrary counts. While genuine count data models can do that, ordered models can only predict outcomes that are actually observed in the sample."*

Winkelmann's critique of the *"ordinal models"* – of which the ***MCL*** is the most representative – is too severe in its first two points, and a bit at odds with his first appreciative statement. Nevertheless, we can only agree with his conclusive comment (ibidem): "*the use of ordered models for count data, and the interpretation of the results, has to proceed with necessary caution. In practice, applications of ordered models to count data are uncommon*" … So rare indeed that Winkelmann did not make reference to any such application, and we, ourselves, could not find one.

Now**,** as far as transportation logistics are concerned, we weighed up the value added by the inclusion and parameterization of *QoS* thresholds which ordinal regression models ordain. Notwithstanding Agresti (2019, p. 170) stated that shift-parameters (specifically, the $\theta_{\mathbb{t}}$-intercepts in the ***MCL*** specification) are usually *"not of interest except for estimating response probabilities"*, here above, we proved they are most relevant to grasp the highly nonlinear variations in such probabilities. Thereby, we demonstrated the instrumentality of these constants generally considered mere *"intercepts"* not even worth commenting: in this respect, Figure **7** is particularly instructive. Correlatively, we proved - from the service provider's viewpoint - that ***time measures had better not be treated* as *cardinal numbers: delivery time must rather be handled as a categorial ordinal variable***. And we believe that what goes for purveyors goes also for their clients, whose ***perceptions*** of delays are likely to be biased upwards. Consequently, count models fall short of accounting for the true nature of durations in the supply chain context. Along those lines, such statistics as mean durations (or delays) can only be misleading: a case in point is the ***UPU (Universal Postal Union)*** who solely displays and expounds evolutions of *"averages"* and *"standard deviations"* of delivery times (2023, pp. 50-51) in their report on the postal sector.

6. Punctuality metrics

The wrap-up of part **5** rationalized the use of the *MCL* model to infer the aggregate *KPIs*:

$$\underline{\hat{\Pi}}(\boldsymbol{O2D},\mathbb{t}) \leq \hat{\Pi}(\boldsymbol{O2D},\mathbb{t}) \leq \overline{\hat{\Pi}}(\boldsymbol{O2D},\mathbb{t}), \text{for } O2D \in \{Co2Cd, Co2Eu, Eu2Cd, Eu2Eu\} \text{ and } \mathbb{t} \in \{1,2,3,4,5,6,7,8,9,10\}.$$

These indicators can indeed be calculated from the *strata-level estimates of probabilities of delivery within a fixed deadline of* $\mathbb{t}$ *periods:* $\hat{\Pi}_{s,\mathbb{t}}$ as explained in sections 3.2 to 3.4. The latter can themselves be predicted for all relevant strata - i.e., non-zero weight, surveyed as well as non-tested ones: *ANS* = 333,792 - from the estimates of the parameters of the *MCL* model, through the *SAS* **P**ostfitting **L**inear **M**odel (**PLM**) procedure (§ ***VI.1*** in Annex **A4**).

6.1. Approaching properties of KPIs estimates

Gauging the imprecision of the *KPIs* is less straightforward. Indeed, the standard errors of the $\hat{\Pi}(\boldsymbol{O2D},\mathbb{t})$, and their bounds**,** depend on the extremely numerous variances of, and covariances between their $\hat{\Pi}_{s,\mathbb{t}}$-components - i.e., $ANS \times (ANS-1)/2 = 55{,}708{,}382{,}736$ - which themselves result from combinations of variances of, and covariances between the estimators of the many (86) parameters - i.e., the 76 coefficients[25]:, and 10 thresholds: $\theta_{\mathbb{t}|B}$ - in equation (15.3). Thus, working out the $[86 \times 86]$-Hessian of the loglikelihood of the sample, inverting that matrix, relying on the Cramér-Rao lower bound to derive the huge variance-covariance matrix of the $\hat{\Pi}_{s,\mathbb{t}}$ and then deduce the standard errors of the *KPI* estimates is a rather demanding process. Therefore, we opted for the ***bootstrapping technique*** which regards the sample on hand $(\mathbb{S})$ as if it were the parent population itself (Efron and Tibshirani,1994). Figure **11** sketches this process.

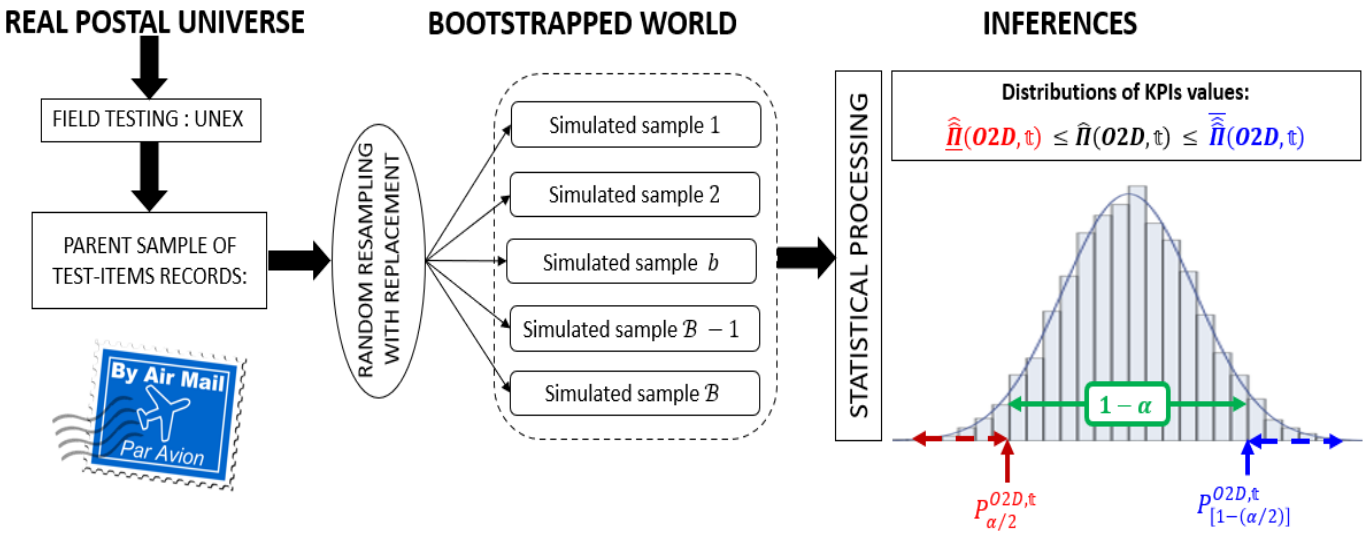


Figure 11. Resampling data to mimic the universe.

Table **11** details it. Annex **A4** explains the code programmed to run it. It infers results for a universe, from estimates compiled by repeating the econometric analysis - as carried out in section **5** - on a sufficiently large number (denoted $\mathcal{B}$) of samples of the same size $(n = 105{,}889)$, generated by random selection - ***with replacement*** - from the available sample of test-records. In the following, only the results obtained for $\mathbb{t} = 5$ are reported. These are what *IPC* calls the ***reliability*** KPIs.

[25] $\delta^{f}_{m(f)/\sigma}$, cf. bottom of the second column of Table **1**: $77 - 1 = 76$, as Ad is no longer considered.

Table 11. Inference algorithm through resampling.

I. Tabulate the records of the n sampled test-items, in the $\mathcal{R}$-matrix, one column per variable, one row per item:

$$\mathcal{R} = \{t_i, Co_i, Uo_i, Zo_i, Sw_i, Fk_i, Wd_i, Pl_i, Cd_i, Ud_i, Zd_i \mid i = 1, \dots, n\},$$

where: $Zo_i = Co_i \cap Uo_i$ and $Zd_i = Cd_i \cap Ud_i$ identify the loci of origin and destination of item i.

II. From $b = 1$ *to* $b = \mathcal{B}$, replicate the following steps: *II.1* to *II.4*

II.1. Draw from the original base sample: $\mathbb{S}$, randomly – with equiprobability $(1/n)$ and ***with replacement*** – n rows from the records' table. This b^{th} bootstrapped sample so extracted from $\mathcal{R}$ is denoted: $\mathcal{R}^{(b)}$ and defined by:

$$\mathcal{R}^{(b)} = \left\{t_{i^{(b)}}, Co_{i^{(b)}}, Uo_{i^{(b)}}, Zo_{i^{(b)}}, \dots, Ud_{i^{(b)}}, Zd_{i^{(b)}}, \mid i^{(b)}\text{: identifies one of the } n \text{ rows picked in } \mathcal{R}\right\}.$$

II.2. Estimate the model fixed parameters $\theta_{\mathbb{t}|B}$ and $\delta^{f}_{m(f)/\sigma}$, and the variances of the random components: $[\varsigma^{O}/\sigma]^2$ and $[\varsigma^{D}/\sigma]^2$, applying the ***GLIMMIX*** procedure on this bootstrapped sample: $\mathcal{R}^{(b)}$.

II.3. From these estimates, predict the on-time delivery probabilities for all of the *ANS* strata (s) forming the full design tree: $\hat{\Pi}^{(b)}_{s|\mathbb{t}}$, from the point-estimates of the parameters of the mean linear predictor function values, to which the local disturbances, simulated by random drawing from normally distributed zero-mean variates: $\sim\mathcal{N}$, are added. Thus,

$$\hat{q}^{(b)}_{s|\mathbb{t}} = -\hat{\theta}^{(b)}_{\mathbb{t}|B} + \left(\sum_{f\in\mathcal{F}}\left(\sum_{m(f)\in\{M^f \mid m(f)\neq b(f)\}}\left[\hat{\delta}^{f}_{m(f)/\sigma}\right]^{(b)} \cdot x^{f}_{s,m(f)}\right)\right) + \left(\sum_{o\in\mathcal{O}}\left[\hat{v}^{O}_{o/\sigma}\right]^{(b)} \cdot Z^{O}_{s,o} + \sum_{d\in\mathcal{D}}\left[\hat{v}^{D}_{d/\sigma}\right]^{(b)} \cdot Z^{D}_{s,d}\right),$$

$$\text{where: } \left[\hat{v}^{O}_{o/\sigma}\right]^{(b)} \sim \mathcal{N}\left(0, \widehat{\varsigma^{O}/\sigma}^{(b)}\right) \text{ and } \left[\hat{v}^{D}_{d/\sigma}\right]^{(b)} \sim \mathcal{N}\left(0, \widehat{\varsigma^{D}/\sigma}^{(b)}\right)$$

$$\Downarrow$$

$$\hat{\Pi}^{(b)}_{s|\mathbb{t}} = 1/\left[1 + e^{-\hat{q}^{(b)}_{s|\mathbb{t}}}\right].$$

II.4. Derive the various estimates of the *KPIs*, from the $\hat{\Pi}^{(b)}_{s|\mathbb{t}}$ using the sets of weights derived from the linear optimization programs specified in Table **4**. File them into the three vectors summing up the b^{th} simulation step:

(a) Lower bounds, using the weights minimizing the $\hat{\Pi}(Co2Cd, \mathbb{t})$:

$$\underline{\hat{\Pi}}(\mathbb{t})^{(b)} := \left\{\underline{\hat{\Pi}}(Eu2Eu, \mathbb{t})^{(b)};\ \underline{\hat{\Pi}}(AT2Eu, \mathbb{t})^{(b)}, \dots, \underline{\hat{\Pi}}(SK2Eu, \mathbb{t})^{(b)}; \underline{\hat{\Pi}}(Eu2AT, \mathbb{t})^{(b)}, \dots, \underline{\hat{\Pi}}(Eu2SK, \mathbb{t})^{(b)}\right\}.$$

(b) Upper bounds, using the weights maximizing the $\hat{\Pi}(Co2Cd, \mathbb{t})$:

$$\overline{\hat{\Pi}}(\mathbb{t})^{(b)} := \left\{\overline{\hat{\Pi}}(Eu2Eu, \mathbb{t})^{(b)};\ \overline{\hat{\Pi}}(AT2Eu, \mathbb{t})^{(b)}, \dots, \overline{\hat{\Pi}}(SK2Eu, \mathbb{t})^{(b)}; \overline{\hat{\Pi}}(Eu2AT, \mathbb{t})^{(b)}, \dots, \overline{\hat{\Pi}}(Eu2SK, \mathbb{t})^{(b)}\right\}.$$

(c) Intermediate estimates, using the *standard weighting base*:

$$\hat{\Pi}(\mathbb{t})^{(b)} := \left\{\hat{\Pi}(Eu2Eu, \mathbb{t})^{(b)};\ \hat{\Pi}(AT2Eu, \mathbb{t})^{(b)}, \dots, \hat{\Pi}(SK2Eu, \mathbb{t})^{(b)}; \hat{\Pi}(Eu2AT, \mathbb{t})^{(b)}, \dots, \hat{\Pi}(Eu2SK, \mathbb{t})^{(b)}\right\}.$$

III. Synthesize each of these set of $\mathcal{B}$ vectors

$$\left\{\underline{\hat{\Pi}}(\mathbb{t})^{(1)}, \dots, \underline{\hat{\Pi}}(\mathbb{t})^{(b)}, \dots, \underline{\hat{\Pi}}(\mathbb{t})^{(\mathcal{B})}\right\}; \left\{\hat{\Pi}(\mathbb{t})^{(1)}, \dots, \hat{\Pi}(\mathbb{t})^{(b)}, \dots, \hat{\Pi}(\mathbb{t})^{(\mathcal{B})}\right\}; \left\{\overline{\hat{\Pi}}(\mathbb{t})^{(1)}, \dots, \overline{\hat{\Pi}}(\mathbb{t})^{(b)}, \dots, \overline{\hat{\Pi}}(\mathbb{t})^{(\mathcal{B})}\right\}.$$

Depict the simulated distributions of the *KPI* estimates, through histograms.
Deduce from these distributions, the $(1-\alpha)\%$-confidence error margins on each of the *KPIs*.

6.2. Robustness of the estimates of the KPIs

The number $\mathcal{B}$ of bootstrapped resamples – identified by (b)-superscripts – required to get dependable point and interval estimates should be high enough to probe the tails of the estimators' distributions. Therefore, given that *"the use of 2000 replicates has been suggested by different sets of authors as reasonable when estimating bootstrap percentile intervals"* (Austin and Leckie, 2020, p. 3196), $\mathcal{B}$ was set equal to **10,000**. To make sure that such a high $\mathcal{B}$ is sufficient, one must first check that the simulated distributions of the *KPI* estimates look *normal*, because they result from maximum-likelihood estimators calibrated on a sample so large that asymptotic properties should hold. This is the case, as evidenced by Figure **12**, for the *Eu2Eu-KPIs*. These four histograms are identically scaled to facilitate comparisons. As expected, they all reveal symmetrical and bell-shaped: gaussian density curves fit them perfectly well. Those in the left panels - depicted for benchmarking purposes - result from simulations where local random variations were neutralized.

The right panel graphs display the distributions of *KPI* values impacted by the spatial heterogeneity. These are much more dispersed than those in the left ones, they are the ones from which to draw the extreme confidence limits:

- the lower bound from the upper rectangle, that of the minima,
- the upper bound from the lower rectangle, that of the maxima, which of course appears to be shifted to the right with respect to the top one, as it should.

Obviously, the uncertainty caused by spatial heterogeneity matters much more than that attributable to the lack of knowledge of the exact weighting schemes.

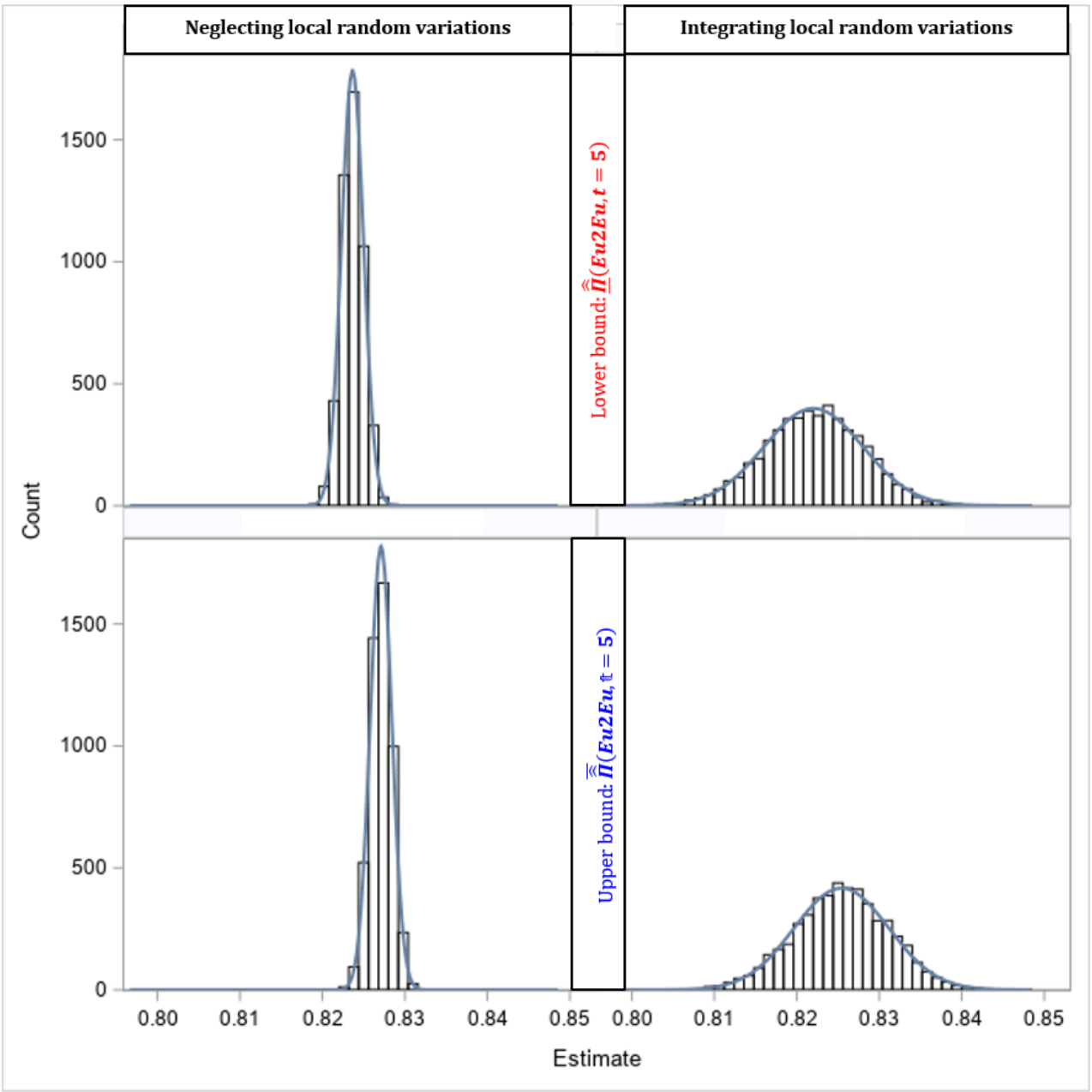


Figure 12. Histograms of *Eu2Eu-KPIs* estimates derived from the 10,000 bootstrap samples.

Figure **12** passes the face-validity tests but just eyeballing it does not fully establish the consistency of the bootstrapping results. These are further assessed by the statistics explained in Table **12**,

Table 12. Two alternative modes of calculation of boundary critical points.

STATISTICS	LOWER BOUND	UPPER BOUND
Mean	$\underline{\hat{\mu}}(O2D,\mathbb{t}) = \frac{1}{\mathcal{B}}.\sum_{b=1}^{b=\mathcal{B}} \underline{\hat{\Pi}}(\boldsymbol{O2D},\mathbb{t})^{(b)}$	$\overline{\hat{\mu}}(O2D,\mathbb{t}) = \frac{1}{\mathcal{B}}.\sum_{b=1}^{b=\mathcal{B}} \overline{\hat{\Pi}}(\boldsymbol{O2D},\mathbb{t})^{(b)}$
Standard deviation	$s\left[\underline{\hat{\mu}}(O2D,\mathbb{t})\right] = \sqrt{\frac{1}{\mathcal{B}-1}.\sum_{b=1}^{b=\mathcal{B}}\left(\underline{\hat{\Pi}}(\boldsymbol{O2D},\mathbb{t})^{(b)} - \underline{\hat{\mu}}(O2D,\mathbb{t})\right)^2}$	$s\left[\overline{\hat{\mu}}(O2D,\mathbb{t})\right] = \sqrt{\frac{1}{\mathcal{B}-1}.\sum_{b=1}^{b=\mathcal{B}}\left(\overline{\hat{\Pi}}(\boldsymbol{O2D},\mathbb{t})^{(b)} - \overline{\hat{\mu}}(O2D,\mathbb{t})\right)^2}$
Normalized confidence limits (*z*-based)	$NCI_{\alpha/2}^{O2D,\mathbb{t}} = \underline{\hat{\mu}}(O2D,\mathbb{t}) - z_{[1-(\alpha/2)]}.\, s\left[\underline{\hat{\mu}}(O2D,\mathbb{t})\right]$	$NCI_{[1-(\alpha/2)]}^{O2D,\mathbb{t}} = \overline{\hat{\mu}}(O2D,\mathbb{t}) + z_{[1-(\alpha/2)]}.\, s\left[\overline{\hat{\mu}}(O2D,\mathbb{t})\right]$
Percentiles	$P_{\alpha/2}^{O2D,\mathbb{t}}$ such that: $P\left(\underline{\hat{\Pi}}(\boldsymbol{O2D},\mathbb{t})^{(b)} \le P_{\alpha/2}^{O2D,\mathbb{t}}\right) = \alpha/2$	$P_{[1-(\alpha/2)]}^{O2D,\mathbb{t}}$ such that: $P\left(\overline{\hat{\Pi}}(\boldsymbol{O2D},\mathbb{t})^{(b)} \ge P_{[1-(\alpha/2)]}^{O2D,\mathbb{t}}\right) = \alpha/2$
Error margins: ***CI*-widths**	$\varepsilon_{1-\alpha}^{O2D,\mathbb{t}} = \begin{cases} \text{Based on percentiles: } \left(P_{[1-(\alpha/2)]}^{O2D,\mathbb{t}} - P_{\alpha/2}^{O2D,\mathbb{t}}\right) \\ \text{Based on the fitted normal distribution: } \left(NCI_{[1-(\alpha/2)]}^{O2D,\mathbb{t}} - NCI_{\alpha/2}^{O2D,\mathbb{t}}\right) \end{cases}$	

One can indeed verify that the limits of the confidence intervals can be

- either, directly determined, non-parametrically, by the relevant percentiles, as in Figure **11**: $\left[P_{\alpha/2}^{O2D,\mathbb{t}}, P_{[1-(\alpha/2)]}^{O2D,\mathbb{t}}\right]$,
- or, obtained by subtracting/adding the value of the standard normal variate, $\tilde{Z}$, corresponding to the $[1-(\alpha/2)]^{th}$ centile: $z_{[1-(\alpha/2)]}$, times the standard deviation from the mean[26]: $\left[NCI_{\alpha/2}^{O2D,\mathbb{t}}, NCI_{[1-(\alpha/2)]}^{O2D,\mathbb{t}}\right]$.

Table **13** confirms that for the usual confidence-levels – $\mathbf{1-\alpha} = 99\%, 95\%$ and 90% – the z-based approach and the percentile method yield rather close values: $NCI_{\alpha/2}^{Eu2Eu,5} \cong P_{\alpha/2}^{Eu2Eu,5}$ and $NCI_{[1-(\alpha/2)]}^{Eu2Eu,5} \cong P_{[1-(\alpha/2)]}^{Eu2Eu,5}$.

So, the most extreme bounds are:

(1) for the lower, from $\underline{\hat{\Pi}}(\boldsymbol{Eu2Eu,5}) \Rightarrow \begin{cases} NCI_{0.5\%}^{Eu2Eu,5} = 80.656\% \text{ or } P_{0.5\%}^{Eu2Eu,5} = 80.616\% \\ NCI_{2.5\%}^{Eu2Eu,5} = 81.025\% \text{ or } P_{2.5\%}^{Eu2Eu,5} = 80.998\% \\ NCI_{5\%}^{Eu2Eu,5} = 81.214\% \text{ or } P_{5\%}^{Eu2Eu,5} = 81.188\% \end{cases}$

(2) for the upper, from $\overline{\hat{\Pi}}(\boldsymbol{Eu2Eu,5}) \Rightarrow \begin{cases} NCI_{99.5\%}^{Eu2Eu,5} = 84.019\% \text{ or } P_{99.5\%}^{Eu2Eu,5} = 83.962\% \\ NCI_{97.5\%}^{Eu2Eu,5} = 83.667\% \text{ or } P_{97.5\%}^{Eu2Eu,5} = 83.631\% \\ NCI_{95\%}^{Eu2Eu,5} = 83.486\% \text{ or } P_{95\%}^{Eu2Eu,5} = 83.462\% \end{cases}$

[26] These are called *"studentized"*, or "bootstrap-*t*", confidence limits, by reference to the *Student's* t-distribution, which can be approximated by the standardized normal, denoted z, when the number of degrees of freedom is large (> 30).

Table 13. Ranges of estimates for *Eu2Eu-KPIs.*

SPATIAL HEREOGENEITY	CONFIDENCE INTERVALS DERIVED FROM THE FITTED NORMAL DISTRIBUTIONS									
	Estimators	Limits of normalized confidence intervals: $NCI_{\alpha/2}^{Eu2Eu,5}, NCI_{[1-(\alpha/2)]}^{Eu2Eu,5}$						Error margins: $\varepsilon_{1-\alpha}^{Eu2Eu,5}$		
		0.5%	99.5%	2.5%	97.5%	5%	95%	99%	95%	90%
Excluded	$\underline{\widehat{\Pi}}(Eu2Eu,5)$	82.017%	82.713%	82.100%	82.630%	82.143%	82.587%	0.696%	0.530%	0.445%
	$\widehat{\Pi}(Eu2Eu,5)$	82.218%	82.902%	82.300%	82.820%	82.342%	82.779%	0.684%	0.521%	0.437%
	$\overline{\widehat{\Pi}}(Eu2Eu,5)$	82.367%	83.053%	82.449%	82.971%	82.491%	82.929%	0.686%	0.522%	0.438%
Included	$\underline{\widehat{\Pi}}(Eu2Eu,5)$	80.656%	83.746%	81.025%	83.377%	81.214%	83.188%	3.090%	2.351%	1.973%
	$\widehat{\Pi}(Eu2Eu,5)$	80.901%	83.889%	81.258%	83.532%	81.441%	83.349%	2.988%	2.273%	1.908%
	$\overline{\widehat{\Pi}}(Eu2Eu,5)$	81.067%	84.019%	81.420%	83.667%	81.601%	83.486%	2.952%	2.246%	1.885%
SPATIAL HEREOGENEITY	DEDUCED FROM THE EMPIRICAL DISTRIBUTIONS OF BOOTSTRAP-ESTIMATES									
	Estimators	Percentiles directly deduced from cumulated histogram: $P_{\alpha/2}^{Eu2Eu,5}, P_{[1-(\alpha/2)]}^{Eu2Eu,5}$						Error margins: $\varepsilon_{1-\alpha}^{Eu2Eu,5}$		
		0.5%	99.5%	2.5%	97.5%	5%	95%	99%	95%	90%
Excluded	$\underline{\widehat{\Pi}}(Eu2Eu,5)$	82.015%	82.710%	82.096%	82.628%	82.142%	82.587%	0.694%	0.531%	0.445%
	$\widehat{\Pi}(Eu2Eu,5)$	82.219%	82.898%	82.296%	82.819%	82.344%	82.777%	0.679%	0.524%	0.433%
	$\overline{\widehat{\Pi}}(Eu2Eu,5)$	82.367%	83.051%	82.449%	82.970%	82.494%	82.927%	0.684%	0.521%	0.433%
Included	$\underline{\widehat{\Pi}}(Eu2Eu,5)$	80.616%	83.719%	80.998%	83.322%	81.188%	83.151%	3.103%	2.325%	1.962%
	$\widehat{\Pi}(Eu2Eu,5)$	80.862%	83.859%	81.241%	83.486%	81.419%	83.322%	2.997%	2.245%	1.903%
	$\overline{\widehat{\Pi}}(Eu2Eu,5)$	81.009%	83.962%	81.393%	83.631%	81.582%	83.462%	2.953%	2.239%	1.880%

And the most extreme error margins are determined by the difference between the highest upper bound on $\overline{\widehat{\Pi}}(\boldsymbol{Eu2Eu,5})$, and the lowest lower bound on $\underline{\widehat{\Pi}}(\boldsymbol{Eu2Eu,5})$. Thus,

$$\varepsilon_{1-\alpha}^{Eu2Eu,5} = \begin{cases} \varepsilon_{99\%}^{Eu2Eu,5} = \left(NCI_{99.5\%}^{Eu2Eu,5} - NCI_{0.5\%}^{Eu2Eu,5}\right) = 3.364\% \text{ or } \left(P_{99.5\%}^{Eu2Eu,5} - P_{0.5\%}^{Eu2Eu,\mathbb{t}}\right) = 3.346\% \\ \varepsilon_{95\%}^{Eu2Eu,5} = \left(NCI_{97.5\%}^{Eu2Eu,5} - NCI_{2.5\%}^{Eu2Eu,5}\right) = 2.641\% \text{ or } \left(P_{97.5\%}^{Eu2Eu,5} - P_{2.5\%}^{Eu2Eu,5}\right) = 2.634\% \\ \varepsilon_{90\%}^{Eu2Eu,5} = \left(NCI_{95\%}^{Eu2Eu,5} - NCI_{5\%}^{Eu2Eu,5}\right) = 2.272\% \text{ or } \left(P_{95\%}^{Eu2Eu,5} - P_{5\%}^{Eu2Eu,5}\right) = 2.274\%. \end{cases}$$

Whatever the way they are arrived at, they seem reliable since both methods converge to practically equal margin-sizes.

Table **14** complements the evidence of reliability provided above for the overall aggregate *Eu2Eu-KPI*. It sums up the regression analyses of

- *means* on *medians,* to check the symmetry of the distributions,
- and of extreme *normalized* confidence limits on corresponding *percentiles:*

$NCI_{\alpha/2}^{O2D,5}$ *on* $P_{\alpha/2}^{O2D5\mathbb{t}}$ and $NCI_{1-(\alpha/2)}^{O2D,5}$ *on* $P_{1-(\alpha/2)}^{O2D,5}$

from both the *outbound* and *inbound* outlooks: $O2D \in \{Co2Eu, Eu2Cd\}$,

and for the 95% and 99% confidence levels.

All coefficients of determination $\left(R^2\text{: rounded to the } 5^{\text{th}} \text{ decimal place}\right)$ are close to *1* indicating a quasi-perfect correlation between the associated dependent and independent variables. Intercepts are very close to ***0***, coefficients neighbor ***1***, which implies quasi-equalities.

Table 14. Additional tests of consistency.

TESTS	$\underline{\widehat{\Pi}}(Co2Eu,5)$			$\overline{\widehat{\Pi}}(Co2Eu,5)$		
	OUTBOUND PERSPECTIVE*: *Co2Eu					
SYMMETRY	***Regression of Means on Medians***					
	Intercept	**Coefficient**	R^2	**Intercept**	**Coefficient**	R^2
Estimate	-2.31E-04	1.00087	1	-0.00018	1.00079	1
p-value	21.27%	0.00%		31.54%	0.00%	
99% Confidence	***Regression of*** $NCI^{Co2Eu,5}_{0.5\%}$ ***on*** $P^{Co2Eu,5}_{0.5\%}$			***Regression of*** $NCI^{Co2Eu,5}_{99.5\%}$ ***on*** $P^{Co2Eu,5}_{99.5\%}$		
	Intercept	**Coefficient**	R^2	**Intercept**	**Coefficient**	R^2
Estimate	0.00130	0.99466	1	-0.00011	0.99692	0.99993
p -value	29.22%	0.00%		93.16%	0.00%	
95% Confidence	***Regression of*** $NCI^{Co2Eu,5}_{2.5\%}$ ***on*** $P^{Co2Eu,5}_{2.5\%}$			***Regression of*** $NCI^{Co2Eu,5}_{97.5\%}$ ***on*** $P^{Co2Eu,5}_{97.5\%}$		
	Intercept	**Coefficient**	R^2	**Intercept**	**Coefficient**	R^2
Estimate	0.00042	0.99764	1	0.00057	0.99766	0.99998
p -value	46.54%	0.00%		37.49%	0.00%	
TESTS	$\underline{\widehat{\Pi}}(Eu2Cd,5)$			$\overline{\widehat{\Pi}}(Eu2Cd,5)$		
	INBOUND PERSPECTIVE*: *Eu2Cd					
SYMMETRY	***Regression of Means on Medians***					
	Intercept	**Coefficient**	R^2	**Intercept**	**Coefficient**	R^2
Estimate	-4.98E-04	1.00126	1	-3.36E-04	1.00107	1
p -value	0.47%	0.00%		8.75%	0.00%	
99% Confidence	***Regression of*** $NCI^{Eu2Cd,5}_{0.5\%}$ ***on*** $P^{Eu2Cd,5}_{0.5\%}$			***Regression of*** $NCI^{Eu2Cd,5}_{99.5\%}$ ***on*** $P^{Eu2Cd,5}_{99.5\%}$		
	Intercept	**Coefficient**	R^2	**Intercept**	**Coefficient**	R^2
Estimate	0.00321	0.99210	0.99993	0.00408	0.99203	0.99994
p -value	0.65%	0.00%		0.08%	0.00%	
95% Confidence	***Regression of*** $NCI^{Eu2Cd,5}_{2.5\%}$ ***on*** $P^{Eu2Cd,5}_{2.5\%}$			***Regression of*** $NCI^{Eu2Cd,5}_{97.5\%}$ ***on*** $P^{Eu2Cd,5}_{97.5\%}$		
	Intercept	**Coefficient**	R^2	**Intercept**	**Coefficient**	R^2
Estimate	0.0014	0.99640	0.99999	0.00177	0.99613	0.99999
p -value	0.26%	0.00%		0.13%	0.00%	

6.3. Contribution of the model-based poststratification

The necessity of poststratification of measurements can be underscored by diagramming the weighted values of the *KPIs* against the corresponding non-weighted sampled in-time proportions. This benchmarking can be visualized no better than by mapping those two sets of estimates on a square Cartesian graph designed to grasp exactly the correction mechanism at work. Figure **13** does that job.

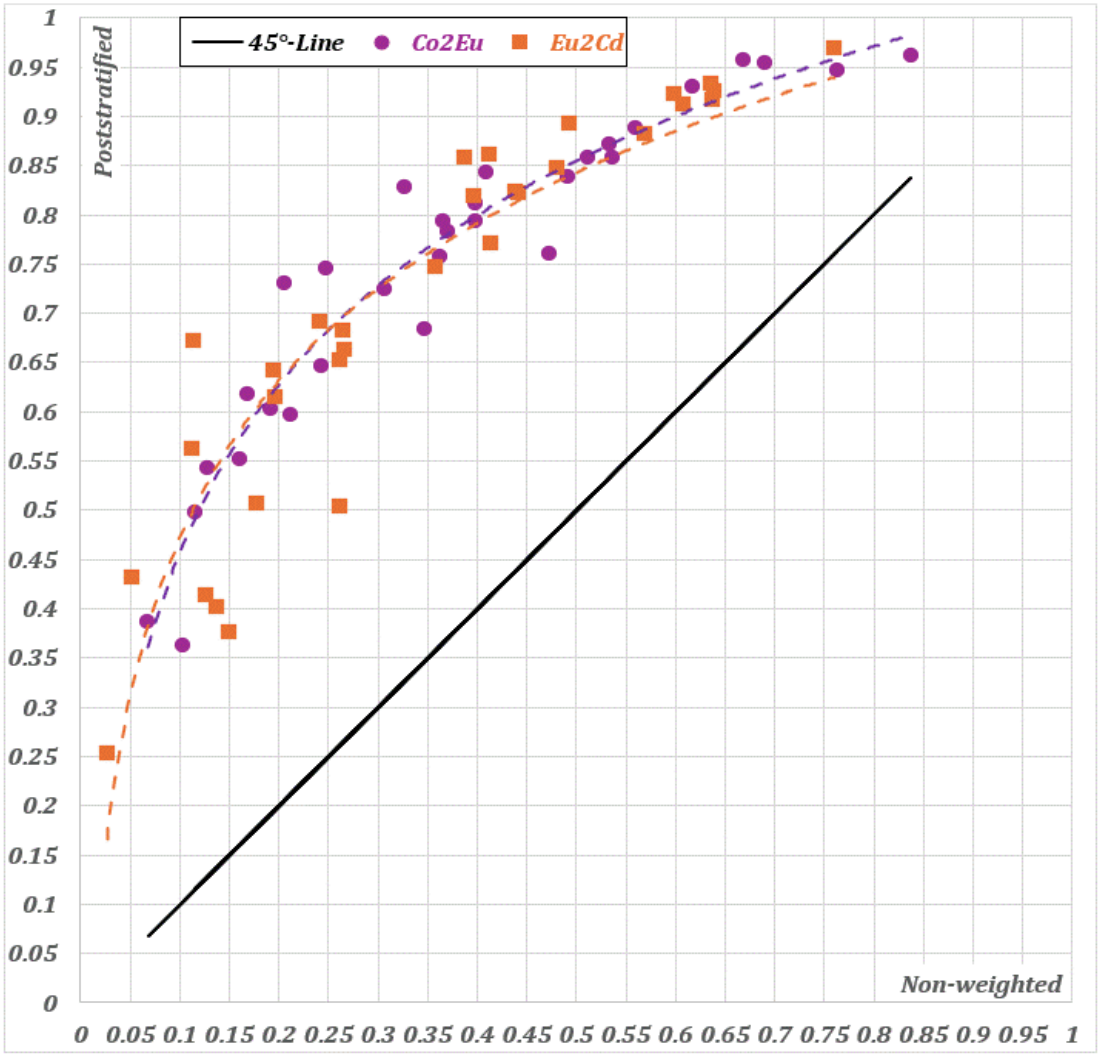


Figure 13. Statistical redressing of sample records into representative *KPIs.*

This mapping shows that

- All of the 62 poststratified *KPIs*, without exception, result from *significant upward adjustments*: all of the *31 purple dots* (*outbound-KPIs*) and *31 orange squares* (*inbound-KPIs*) stand well above the 45°-line.
- These two plots follow roughly the same logarithmic trend: the higher the performance, the nearer the *inbound logistics ratings* get to the *outbound ones*, the less drastic the rectification of the unweighted estimates. Indeed, top operators' achievements are more uniform.

The scale and direction of the poststratification effects are due to the following facts:

(1) The minimum subsample size constraints, imposed by *CEN*, lead to test disproportionally more the small-volume routes, which happen to score very low on the quality-of-service.

(2) Two *franking* modes (*Fk*: metered, prepaid) and one *induction place* (*Pl*: pickup), characteristics of business mail[27], which facilitate envelope handling, are undertested because market research contractors find it difficult to recruit and manage business panelists (company employees).

[27] Note that *metering* is not the franking mode exclusively reserved for business customers because in some post-offices, envelopes inducted over the counters are franked by the staff via metering machines instead of affixing stamps.

6.4. Towards more operational diagnoses

Beware not to be fooled by the big picture that emerges from Figure **12** and Table **13**: the relatively satisfactory, and fairly accurate, overall score:

$$\widehat{\Pi}(\boldsymbol{Eu2Eu},\mathbf{5}) = \mathbf{82.395\%}\begin{cases} -\mathbf{1.779}\ \% \ \text{(from: } P_{0.5\%}^{Eu2Eu,5}\text{)} \\ +\mathbf{1.624\%}\ \text{(from: } NCI_{99.5\%}^{Eu2Eu,5}\text{)} \end{cases}\textbf{, with 99\% confidence,}$$

hides quite disparate realities. Figure **14** proves that, for many countries, achievements at both ends are far from equal.

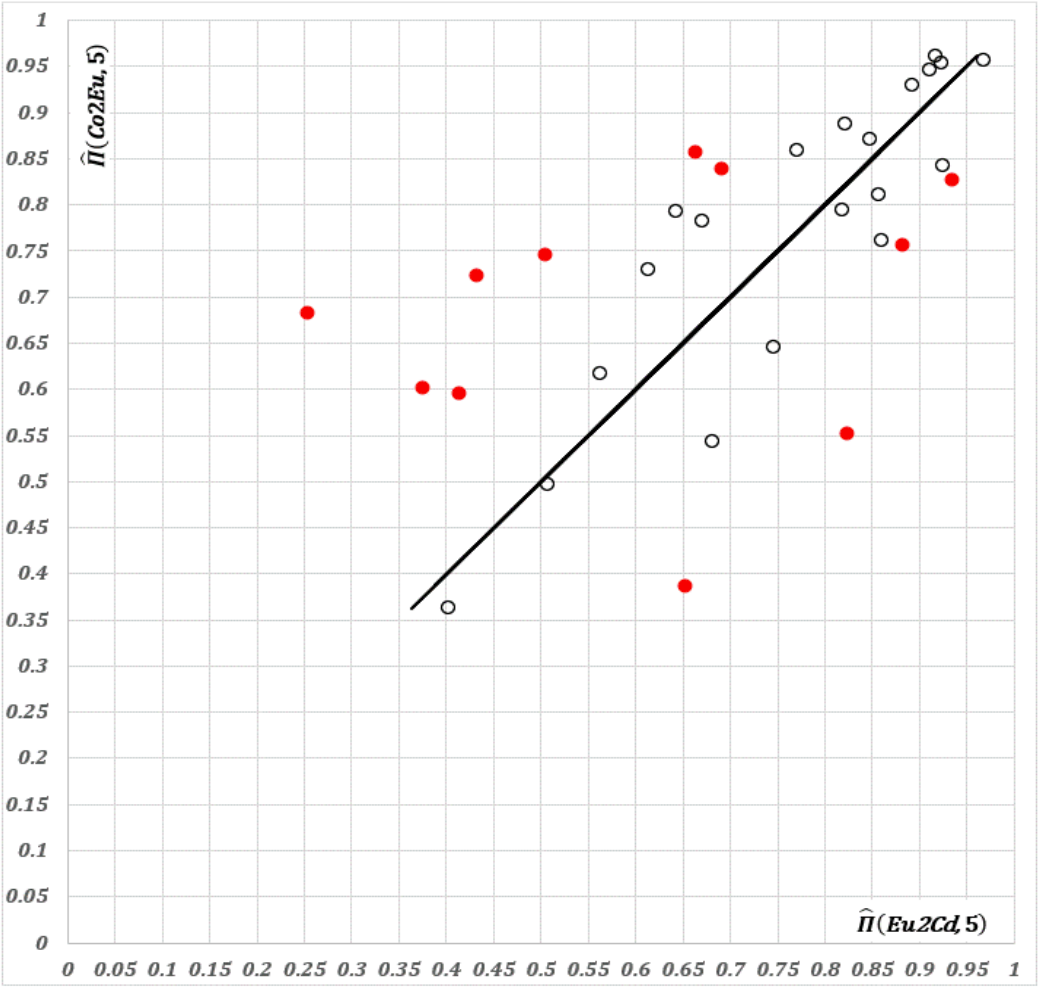


Figure 14. Relationship between outbound and inbound punctuality, for $Co \equiv Cd$.

First of all, the estimates of the in-time delivery probabilities looked at from the outbound (i.e., $Co2Eu$) need be differentiated from those examined from the inbound-side (i.e. the $Eu2Cd$-ones). So, it's vital to ponder both perspectives! Secondly, the scatterplot coupling these two facets of the operations in the 31 countries is funnel-shaped: its dispersion narrows down progressively from the bottom-left up to the top-right. Criteria values get closer and closer to one another as the chances of successful delivery within the deadline improve. This is consistent with the convergence observed in Figure **13**. Posts participating in operations for which the gap is wide (e.g., those pointed by red dots) learn a lot just by internally benchmarking their outbound- against their inbound-processes and by reviewing their chaining. Of course, they may benefit even more from winners as long as they are ready to share their experience, which *IPC* encourages and facilitates through various platforms of exchanges, its main *raison d'être*.

Figure **15** further documents the variance lying under the rather narrow, and high, estimate-bracket of the overall global *Eu2Eu-KPI*. Only the vertical axis is graduated, as the abscissas are meaningless. It displays confidence intervals, in ascending order of *SWB*-weighted mean predicted probability. Limits of these fences are conservative:

$$\textbf{MIN L-99\%} = \boldsymbol{min}\left\{P_{0.5\%}^{O2D,5}, NCI_{0.5\%}^{O2D,5}\right\} \text{ and } \textbf{MAX U-99\%} = \boldsymbol{max}\left\{P_{99.5\%}^{O2D,5}, NCI_{99.5\%}^{O2D,5}\right\}.$$

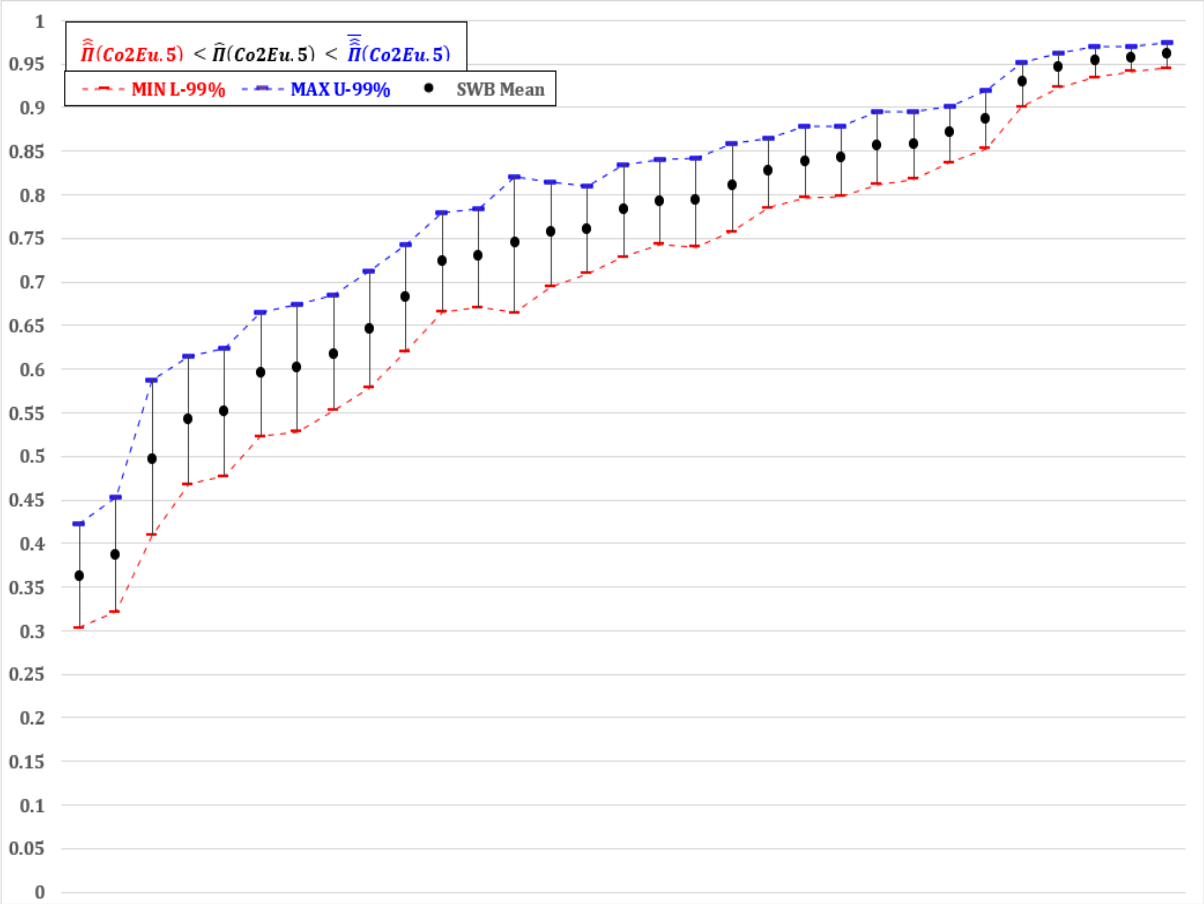


Figure 15.a Point-estimates and 99%-confidence bounds of probabilities of *Co2Eu* delivery within 5 days.

Although the amplitude of their scaling is such that they may look symmetric they are not because part of the uncertainty comes from the weighting schemes (cf. *3.4*, here above). Most striking are the disparities between countries, which for business confidentiality's sake cannot be singled out. Indeed, there is a long way

- from the great depths:
  Figure **15.a**: $[36.26\% - 5.89\% = \mathbf{30.37\%}] < \hat{\boldsymbol{\Pi}}(\boldsymbol{Co2Eu}, \mathbf{5}) = \mathbf{36.26\%} < [36.26\% + 6.06\% = \mathbf{42.32\%}]$
  Figure **15.b**: $[25.35\% - 5.56\% = \mathbf{19.79\%}] < \hat{\boldsymbol{\Pi}}(\boldsymbol{Eu2Cd}, \mathbf{5}) = \mathbf{25.35\%} < [25.35\% + 6.12\% = \mathbf{31.47\%}]$
- up to the high peaks:
  Figure **15.a**: $[96.14\% - 1.64\% = \mathbf{94.50\%}] < \hat{\boldsymbol{\Pi}}(\boldsymbol{Co2Eu}, \mathbf{5}) = \mathbf{96.14\%} < [96.14\% + 1.32\% = \mathbf{97.46\%}]$
  Figure **15.b**: $[96.82\% - 1.24\% = \mathbf{95.58\%}] < \hat{\boldsymbol{\Pi}}(\boldsymbol{Eu2Cd}, \mathbf{5}) = \mathbf{96.82\%} < [96.82\% + 1.05\% = \mathbf{97.87\%}]$.

Not too unexpectedly, their widths almost match the uncertainty about the outcome of the logistic process reflected by the probability of in-time delivery value: the closer it is to 0.5, the wider the confidence interval.

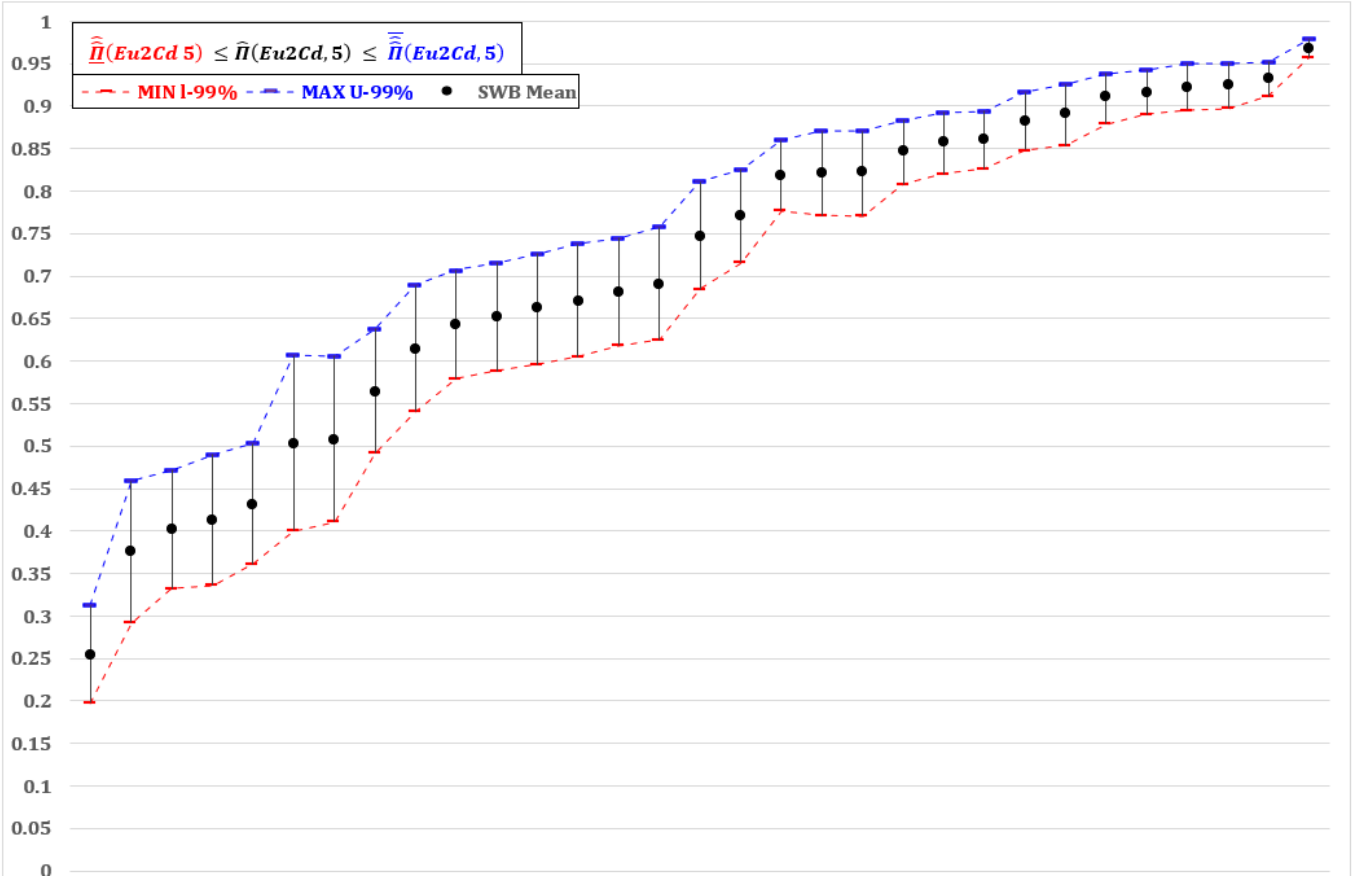


Figure 15.b Point-estimates and 99%-confidence bounds of probabilities of $Eu2Cd$ delivery within 5 days.

6.5. Possibility to reduce the bootstrapping task

For computational efficiency, one should minimize the number of bootstrapped samples needed to infer dependable confidence intervals. $\mathcal{B} = 10{,}000$ proved to be sufficient … But wouldn't 2,000, or even less, be enough? To answer this pragmatic but critical, question, the sensitivity of the estimates of the confidence-interval limits $(CILs)$ to lowering $\mathcal{B}$ should be tested. Their robustness can efficiently be assessed by probing the values arrived at when they are applied to a subset, $\mathcal{s}$ of size: $\mathcal{b} = |\mathcal{s}| < \mathcal{B}$, randomly selected without replacement from the entire set comprised of those derived from the $\mathcal{B}$ bootstrapped samples. The accuracy of the estimates of the $CILs$, reviewed KPI per KPI, can then be measured by the ratios of each of these estimates inferred from such a subset of $\mathcal{b}$ samples: $\widehat{CIL}(\mathcal{s}|\mathcal{b})$, to that derived from the whole set of $\mathcal{B}$ bootstrapped samples: $\widehat{CIL}(\mathcal{B})$:

$$\mathbb{r}_{\mathcal{s}|\mathcal{b}}(CIL) = \frac{\widehat{CIL}(\mathcal{s}|\mathcal{b})}{\widehat{CIL}(\mathcal{B})} \text{ for } CIL \in \{P_{0.5\%}^{O2D,5}, P_{99.5\%}^{O2D,5}, P_{2.5\%}^{O2D,5}, P_{97.5\%}^{O2D,5}, NCI_{0.5\%}^{O2D,5}, NCI_{99.5\%}^{O2D,5}, NCI_{2.5\%}^{O2D,5}, NCI_{97.5\%}^{O2D,5}\},$$

$$\mathcal{s} \in \{1, 2, \ldots, 1\,000\} \text{ and } \mathcal{b} \in \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}.$$

Of course, to learn meaningful lessons from such indicators, many different subsets $\mathcal{s}$ of size: $\mathcal{b}$ need to be drawn, corresponding ratio values be calculated and the distribution of these values analyzed. In fact, the more concentrated this distribution turns out to be around 1, the more reliable the estimate of the CIL is. Hereafter, we base our diagnosis on 1,000 subsamples of estimates per CIL and per KPI. The scope of this investigation was restrained to three $KPIs$: $Eu2Eu$, $Co2Eu$ for a single Co origin country and $Eu2Cd$ for just one Cd destination country. We picked the Co (Cd) country exhibiting the most uncertain level of outbound (inbound) performance: i.e., the one evincing the largest error margins in Figure **15.a** (**15.b**), whose KPI-estimates are most imprecise.

Figures **16.a** and **16.b** sum up, for the *Eu2Eu-KPI*, the distributions – one per subset-size $\mathscr{b}$ – of the 1,000 $\mathbb{r}_{\mathcal{s}|\mathscr{b}}$ ratio-values relevant to the 99%-confidence ***lower*** limits derived from percentiles and studentized statistics, respectively. Analogously, Figures **17.a** and **17.b** sum up the distributions – one per subset-size $\mathscr{b}$ – of the 1,000 $\mathbb{r}_{\mathcal{s}|\mathscr{b}}$ ratio-values relevant to the 99%-confidence ***upper*** limits of the *Eu2Eu-KPI* derived from percentiles and studentized stats, respectively.

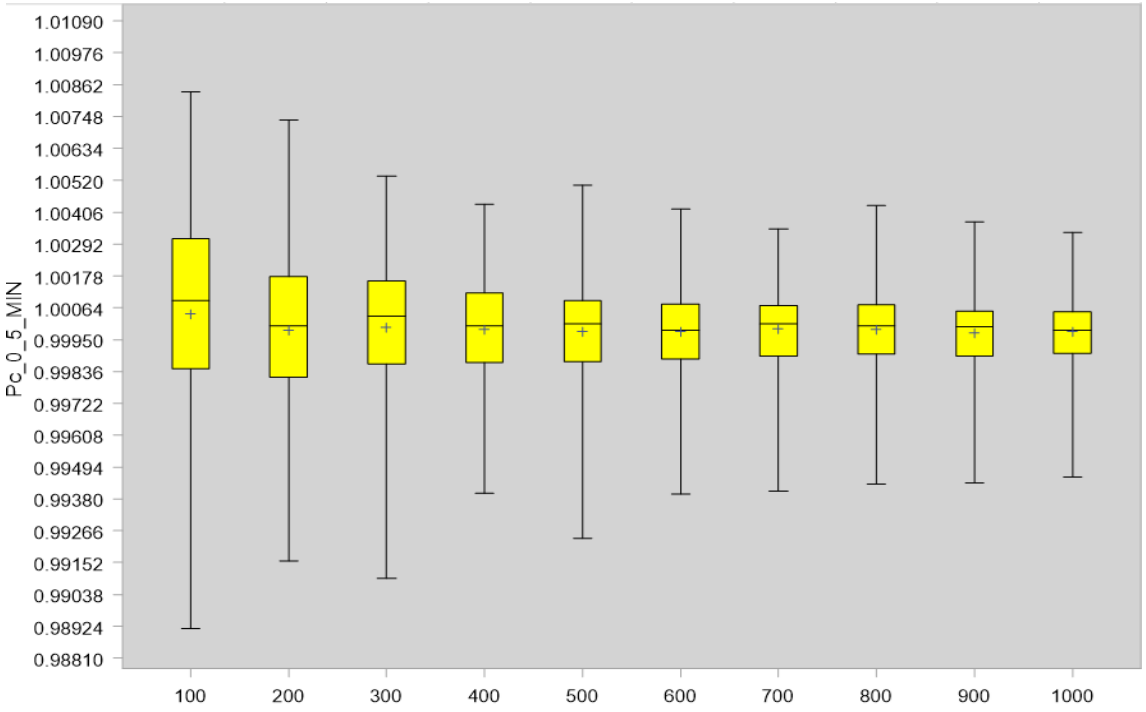


Figure 16.a Distributions of 1,000 values of the ratio related to: $CIL = P_{0.5\%}^{Eu2Eu,5}$.

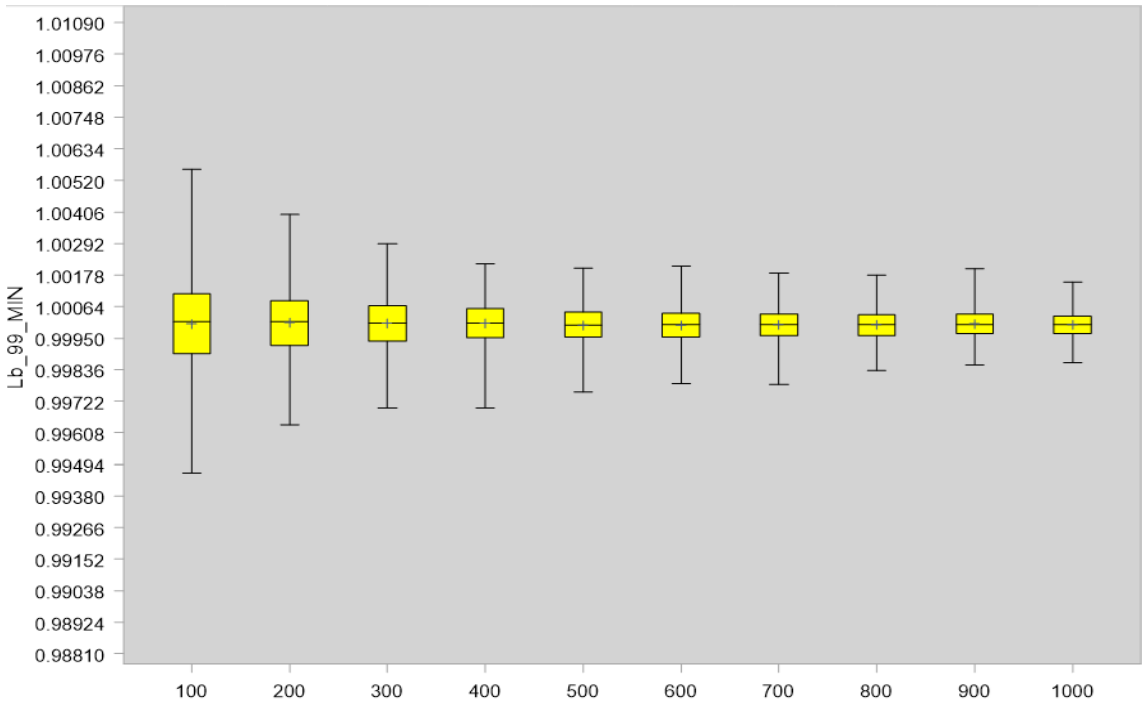


Figure 16.b Distributions of values of the ratio related to: $CIL = NCI_{0.5\%}^{Eu2Eu,5}$.

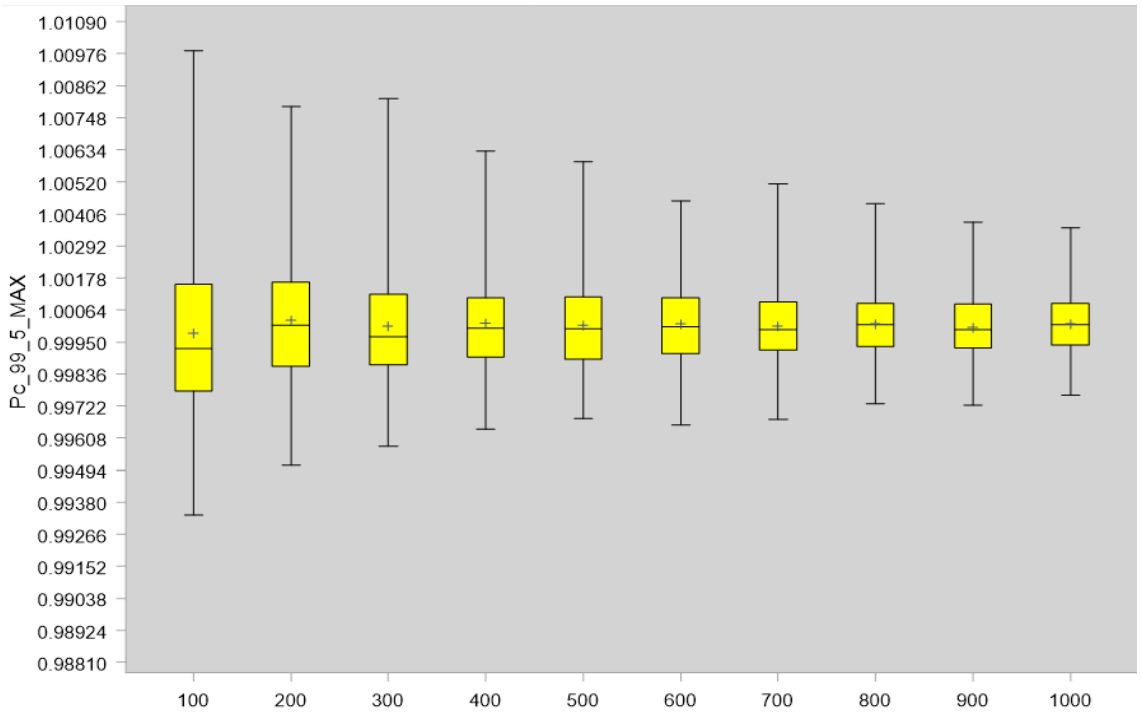


Figure 17.a Distributions of values of the ratio related to: $CIL = P_{99.5\%}^{Eu2Eu,5}$.

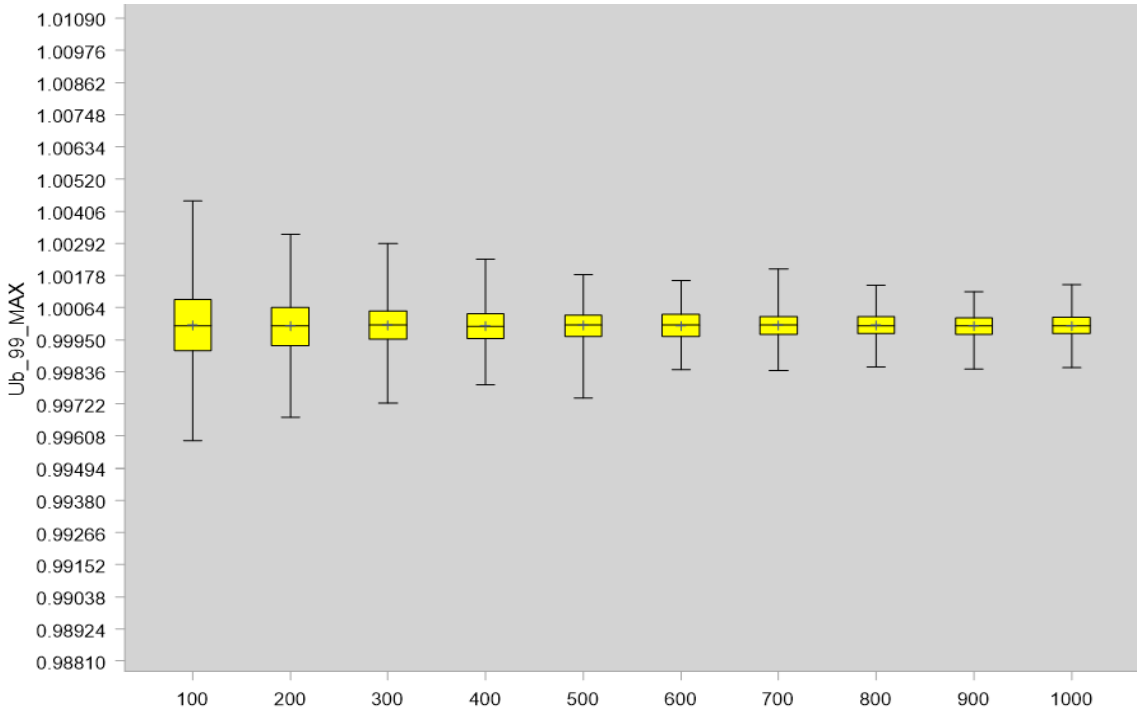


Figure 17.b Distributions of values of the ratio related to: $CIL = NCI_{99.5\%}^{Eu2Eu,5}$.

All four Figures, 16.a to 17.b, have been identically scaled to facilitate comparisons. These box-plotted histograms evince the same pattern:

(a) Their means and medians hardly depart from 1.

(b) As expected, the larger $\mathscr{b}$ is, the smaller their dispersion.

(c) Yet, ratio distributions of percentile-based *CILs* are less concentrated and more asymmetric than those of the studentized ones and more markedly so the lower $\mathscr{b}$ is.

As similar patterns are observed for the 3 *KPIs* and all of the 8 *CILs*, we don't include the other 22 (6+ 2x8) graphs of paralleled boxplots[28]. Rather, to provide a helicopter and comprehensive view on the sensitivity of *CIL* estimates to the bootstrapping rate ($\mathscr{b}$), we plotted the *KPI*-specific links between on the one hand, the ***ranges*** of the ratio-values:

$$\text{Range of } \mathbb{r}_{s|\mathscr{b}}(CIL) = \left[\max_{s}\{\mathbb{r}_{s|\mathscr{b}}(CIL)\} - \min_{s}\{\mathbb{r}_{s|\mathscr{b}}(CIL)\}\right],$$

and on the other, the bootstrapping rate. These synthetic graphs are displayed in Figures **18.a** to **18.c**. (note that the scale bounds of the vertical axis of Figure **18.a** are more than ten times inferior to those of Figures **18.b** and **18.c**). All three Figures consistently confirm observations *(a)* through *(c)* above. In addition, they establish the ranking of the CIL estimates in increasing order of accuracy:

$$P^{O2D,5}_{0.5\%} \prec P^{O2D,5}_{99.5\%} \prec P^{O2D,5}_{2.5\%} \prec P^{O2D,5}_{97.5\%} \prec NCI^{O2D,5}_{0.5\%} \prec NCI^{O2D,5}_{99.5\%} \prec NCI^{O2D,5}_{2.5\%} \prec NCI^{O2D,5}_{97.5\%}.$$

They also show that the precision of all *KPI* estimates had better be assessed by the studentized $CILs$:

$$\left[NCI^{O2D,\mathbb{t}}_{\alpha/2}, NCI^{O2D,\mathbb{t}}_{[1-(\alpha/2)]}\right],$$

than by the percentiles of their empirical distribution. It should be noted, however, that the validity of the results of this second stage – which involves *nested inner* subsampling without replacement from the finite population of 10,000 outer bootstrap estimates, performed here – depends heavily on the representativeness of these first-stage bootstrap estimates, which we believe we have ascertained in *6.2.*

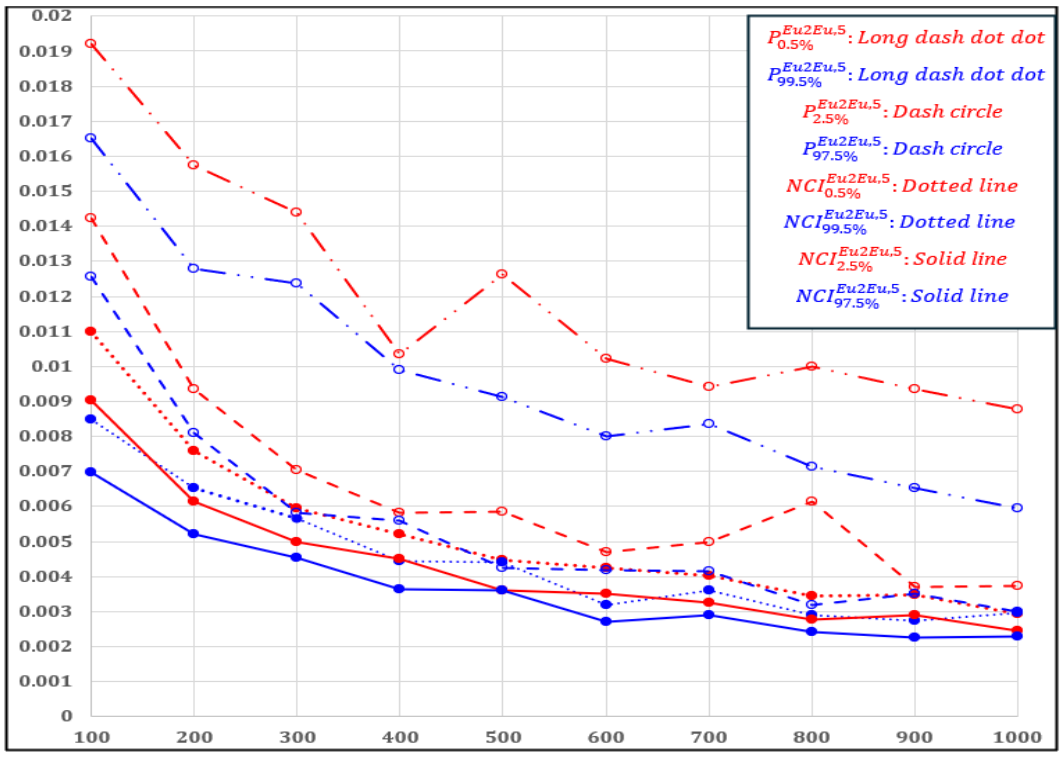


Figure 18.a Ranges of values of the $\mathbb{r}_{s|\mathscr{b}}(CIL)$-ratios for *Eu2Eu*-KPI, as a function of $\mathscr{b}$.

[28] This benchmarking approach has been implemented by means of two complementary *SAS* codes, copies of which can be obtained by contacting the first author, who can also deliver the full set of 24 (3x8) graphs upon requests.

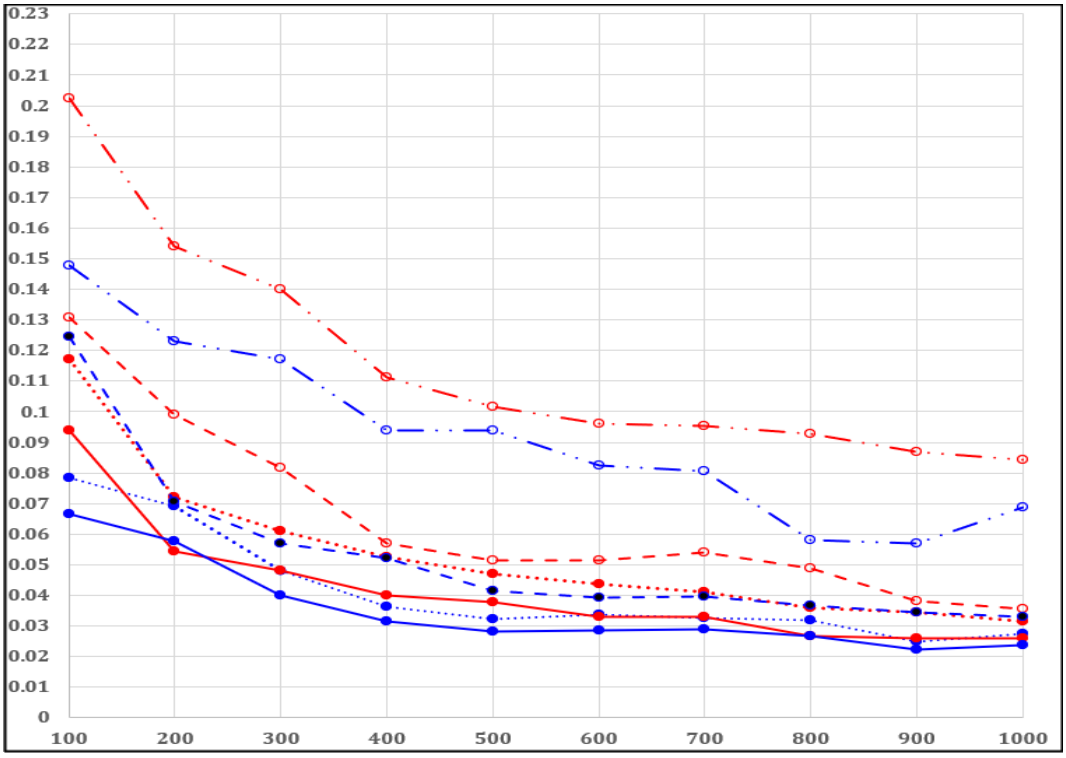


Figure 18.b Ranges of values of the $\mathbb{r}_{s|\mathscr{b}}(CIL)$-ratios for *Co2Eu*-KPI, as a function of $\mathscr{b}$.

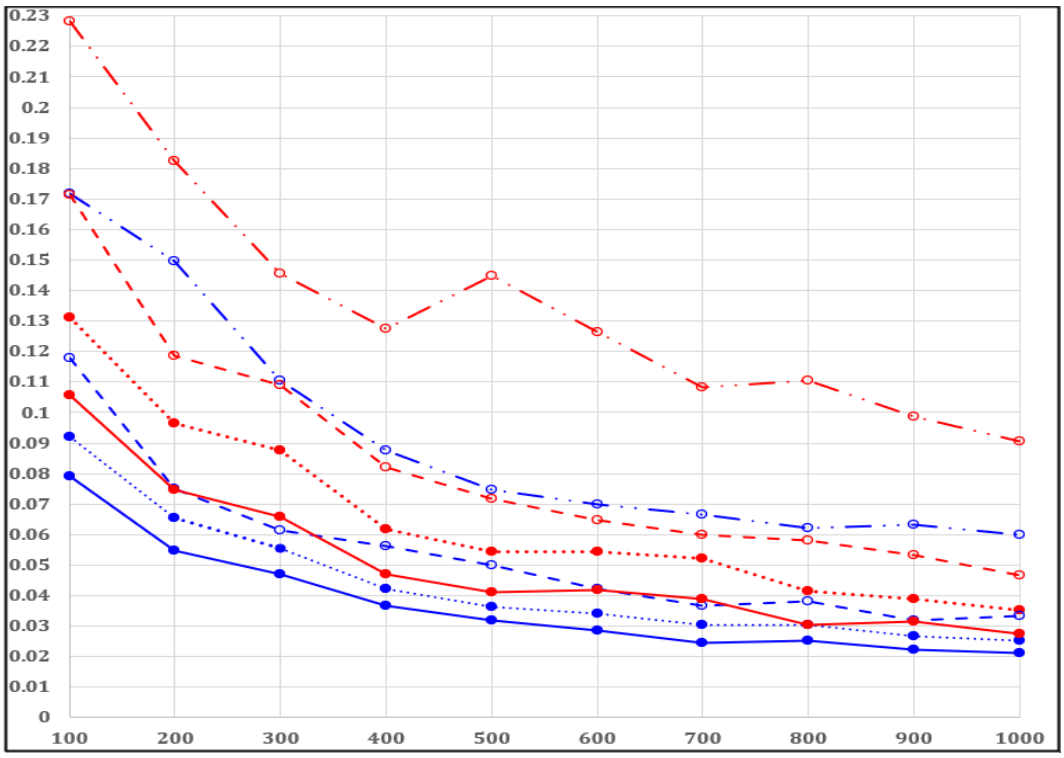


Figure 18.c Ranges of values of the $\mathbb{r}_{s|\mathscr{b}}(CIL)$-ratios for *Eu2Cd*-KPI, as a function of $\mathscr{b}$.

The marked trends towards the horizontal of the four lower curves, highlighted by Figures **18.a** to **18.c**, reveal that to evaluate the precision of the CIL estimates of the overall *Eu2Eu* and *Co2Eu* and *Eu2Cd KPIs*,

- at the 95% confidence level (the two lowest curves), 600 bootstrap resamples are sufficient,
- at the 99% confidence level (the 3rd and 4th curves from bottom), 1000 bootstrap resamples are recommended.

7. Conclusions

The integrative methodology finetuned in this article is implementable not only in the delivery services' sector but to many other cases of failures to fully cover stratified universes, such as for election polls and market-research studies. In fact, it can be usefully adapted in all instances where the number of classification criteria of the population is large but gets practically restricted to no more than a few (most often, age, gender and social class), while they could include additional ones more directly germane to the survey's purpose (e.g., behavioural factors such as political orientation, purchasing power, equipment, …), yet are ignored because of lack of information on strata allocation.

Thus, given its applicability potential, the approach fostered here deserves to be validated. Benchmarking it against alternatives is enlightening:

- *Alternative 1:* Strictly stick to the available disproportionate sample, renouncing any adjustment that could restore proportionality. Such a last resort solution can only lead to biased values, badly so for the Posts: Figure **13** has illustrated how underestimated probabilities of in-time delivery are by the corresponding unredressed proportions.
- *Alternative 2:* Prune the number of mail characteristics to the most discriminant ones and pool their less differentiating modes, as much as possible, to hopefully cut down the number of strata to make sure all are sufficiently sampled. In this respect, ***CEN*** prescribes to take into account *"only those … that prove to be discriminant"* (2020, p. 20): i.e., confirmed so by *"quick-check of significance"* (2020, § **G.1.2.2**, pp. 69-70). To do so, ***CEN*** advises to run t-tests of pairwise differences between means of *"transit times"* of items differentiated by two of the modes of the single characteristic whose effect is analysed independently from others[29], and to consider that this difference is *significantly discriminant* only if the p-value associated with the t-stat exceeds the two-tail sacrosanct 5% α-threshold. Many application-oriented statisticians from all domains have refuted this blind selection process[30] because it is likely to induce conclusions of insignificance of differences which may be practically meaningful, just because of the imprecision of estimates: e.g., Bultez et al. (2022) named *"dichotomania"* such overreliance on the mechanic benchmarking of the p-value against the α-threshold[31]. Bultez and Herrmann (2025) discussed examples of such a tendency. This is why paragraphs 5.3.1-5.3.2, above, rationalize the chosen option of neither eliminating any of the potential predictors, nor grouping modes.

7.1. Main limitations

7.1.1. Unknown extrapolability

The real test of the approach should be to assess the predictive power of the model, which is impossible when many strata are not surveyed or are barely sampled because then there is no reliable counterfactual evidence. As surrogates, Monte Carlo experiments could be programmed, but the results of these computer-lab tests would be biased in favor of the model which would drive the simulation. Thus, the model external validity cannot be established.

[29] These one-factor-two-mode-at-a-time tests are biased because ignored effects of the omitted covariates get picked up by the variable focused on, to the extent of the collinearity between this and the others.

[30] Hulbert et al.'s (2019, pp. 354-355) historical review of the anti-threshold movement starts as early as in 1960.

[31] Moreover, confusions arise because p is often misinterpreted as the probability of the null hypothesis. Worse, $1-p$ is mistaken for the *"level of significance"*, as ***CEN*** believes (op. cit., p. 69: comment on top of Table **G.1**).

7.1.2. Omitted covariates

No matter how comprehensive a model may be, it will always remain incomplete because its very purpose is to simplify a complex reality. Here, factors such as:

- seasonality: extreme peaks in the volumes of mail exchanged during certain periods of the year (such as Christmas/Nex Year) are likely to affect logistics performance,
- interactions: certain combinations of mailings' features - e.g., metered-picked up, may be more easily handled in certain countries' outbound-facilities - might have to be incorporated … But the number of parameters is already very large. Adding more would cause computational issues: prohibitive running time, non-convergence, lack of memory.

Therefore, the introductory part of section 4 invoked Zellner's reinterpretation of the *KISS* principle.

7.1.3. Bootstrapping techniques

The basic bootstrapping with replacement that we resorted to in order to derive confidence intervals for *KPI* estimates (Table 11) is extremely CPU-time consuming. Alternative, more advanced methods may be less computationally intensive, but reviewing and testing them falls outside our scope.

7.2. Future research

Addressing the shortcomings identified in *7.1* would open up new investigation tracks. Yet, other paths might be worth exploring.

7.2.1. Contrasting in-time versus late

The proportional odds hypothesis upon which the multinomial cumulative regression rests is questionable (refer back to § 4.2.3), hence, separate binomial models, one per deadline - contrasting various *"in time"* versus *"late"* events -, could be parameterized as follows:

- *fast* if $\mathbb{t} \leq 2$ versus *low speed* if $\mathbb{t} \geq 3$,
- *rapid* if $\mathbb{t} \leq 3$ versus *slow* if $\mathbb{t} \geq 4$, the distinction ***IPC*** has picked to run stepwise multivariate discriminant analyses to simplify the stratification (*Alternative 2*, alluded to here above),
- *reliable* if $\mathbb{t} \leq 5$ versus *undependable* if $\mathbb{t} \geq 6$.

Which combination of such variants would be best and to what extent would it outclass the multinomial cumulative remain open questions.

7.2.2. Model blending real-world and test data

UNEX™-CEN perspective is holistic in that it covers the entire shipment and delivery process - for the full end-to-end walk of a test-letter all the way through the postal pipeline - and is only interested in measuring the total transit time. Though, systematic ***RFID*** tracking makes it possible to break down each test item's journey into its outbound stretch in the country of origin (leg 1), its cross-border transport (leg 2), and its inbound stretch in the country of destination (leg 3). Moreover, for real-time control of operations, all real mail flows get scanned at the sending Post's international departure center in the country of origin and at the receiving Post's international arrival hub in the country of destination. Therefore, the representativeness of the test-items' course through leg 2 could be truly evaluated. Also, more importantly, an integrative model built to handle both real-world and test data would render sampling intercountry logistical links unnecessary. Then, testing would be confined to the domestic legs 1 and 3 treated as independent stages. In turn, this would simplify the stratification design and sampling plan.

7.2.3. Towards a truly predictive model

While the model currently facilitates post-factum analytics to guide remedial actions, a more consolidative version of it - exploiting both sources of data - could flag up in-process items at risk and Posts could take specific measures so that the much-dreaded delays forecasted could be prevented.

7.3. Originality

Despite its shortcomings, our paper is the first to tackle

- the uncertainty about weighting schemes, which are only partially known: i.e., marginally, per characteristic, and not jointly,
- a system's quantitative response - i.e., delivery time, by nature a *ratio-scaled measurement*[32] - as a categorical ordinal measure and consequently, better appraise what difference a day makes.

Funding Statement

This research received no external funding.

Acknowledgments

We sincerely thank Aude DEVILLE (Institute of Business Administration, Université Côte d'Azur), Christine DI MARTINELLY (IÉSEG School of Management, Lille-Paris), Bart JOURQUIN (Louvain School of Management, Mons), Fouad RIANE (Centrale, Engineering school, Casablanca), Jacques THEPOT (University of Strasbourg), and three anonymous reviewers, for their constructive comments and suggestions on earlier drafts. We would also like to recognize the efficiency of the Editorial Board work. The authors take sole responsibility for any remaining errors.

Conflict of interest

The authors claim that the manuscript is completely original. The authors also declare no conflict of interest.

Appendix

A.1 Program designed to calculate the bounds on point-estimates of country-to-country *KPIs*: cf. subsection *3.4*.

```
/* I. Input: real mail weights of factors' modes, standard weighting schemes, predicted probabilities of delivery by  𝕥  */
/* I.1. Real mail weights of factors' modes:  Ω_m(f|C)  →  Omega_mf       */
DATA RMW; INFILE 'C:\... ... \Pooled_RMW.csv' DLM="," DSD;
        INPUT Factor $ Mode $ Country $ Omega_mf; IF Omega_mf=0 THEN DELETE;
RUN;
/* I.2. Standard Weighting Schemes:  ω_s|Co2Cd  →  SWB_s */
DATA SWB; INFILE 'C:\... ... \Weighting_schemes.csv' DLM="," DSD;
       INPUT Fk $ Pl $ Sw $ Uo $ Wd $ Co $ Ud $ Cd $ SWB_s Co2Cd $;
RUN;
PROC SORT DATA=SWB; BY Co2Cd Fk Pl Sw Ud Uo Wd; RUN;
/* I.3. Counting and numbering of Co2Cd-paths */
PROC SORT DATA=SWB NODUPKEYS OUT=Paths; BY Co2Cd; RUN;
DATA Paths; SET Paths; Path=_N_; KEEP Co2Cd Path; RUN;
DATA _NULL_; SET Paths; END=LAST; IF LAST THEN CALL SYMPUT("N_paths ",LEFT(_N_)); RUN;
DATA SWB; MERGE Paths SWB; BY Co2Cd; RUN;
```

[32] On measurement scales, see: Stevens (1946).

```
/* I.4. Strata-specific predicted probabilities of delivery within 𝕥-days: Π̂_{s,𝕥} → SPD_by_T */
DATA Performance_D5; INFILE 'C:\... ... \ Predictions_t_5.csv' DLM="," DSD;
      INPUT Co $ Cd $ Uo $ Ud $ Sw $ Fk $ Pl $ W_Eu2Eu Wd $ SPD_by_T;
      /* Weighting of the Π̂_{s,𝕥}: ω**_{s|o2d} → W_Eu2Eu */ W_Perf = SPD_by_T*W_Eu2Eu;
RUN;
PROC SORT DATA =Performance_D5; BY Co Cd Fk Pl Sw Ud Uo Wd; RUN;
/* I.5. Benchmark: Overall SWB Eu2Eu predicted probability of delivery by 𝕥: Π̂(Eu2Eu, 𝕥) →Eu2Eu_PD_by_T */
PROC MEANS DATA =Performance_D5 NOPRINT; VAR W_Perf; OUTPUT OUT =KPI_5 SUM=Eu2Eu_PD_by_T; RUN;

/* II. Routine optimizing the point-estimates of the in-time delivery probabilities in-time delivery:
       Π̲̂(Co2Cd, 𝕥) and Π̂̄(Co2Cd, 𝕥) defined in Table 4, consistent with real weights of factors' modes.
Concomitant determination of weighting vectors maximizing/minimizing those point-estimates: w̲_{Co2Cd}, w̄_{Co2Cd}
Called by the %Execution-macro, see below: V */
%MACRO Optimization;
/* The OPTMODEL language enables one to build and solve optimization models: part of the SAS/OR software */
      PROC OPTMODEL;
    /* II.1. Declaration of sets of strings, and matrix & vectors of parameters,
       necessary for the specification of the conditional weights of factors' modes */
          SET Strata;
          SET <str> mfs; /* m(f|C) → mfs: factors' modes */
    /* M_mfs_id: matrix of identifiers factors' modes charactering strata of items */
          NUMBER M_mfs_id {Strata,mfs};
          NUMBER Mf_W {mfs}; /* Mf_W: vector of modes' weights */
    /* Vector of strata-specific predicted probabilities of intime delivery: Π̂_{s,𝕥} → pi_s_t */
          NUMBER pi_s_t {Strata};
    /* II.2. Inputs of constants through statements of the form:
       "READ DATA SAS-data-set INTO set-name =[key-columns]"
       Such a statement reads data from a SAS-datasets into parameters' matrix and vectors' locations.
       Arguments are:
       - SAS-data-set specifies the input data set name;
       - set-name: set vector in which to save the set of observations read from the input data set;
       - key-columns: provide the index values for array: destinations,
       - columns: specify the data values to read and the destination locations.
    /* II.2.1. Importing names of binary variables identifying the factors' modes from the file
       specified just after the DATA-keyword (i.e., I_MF), to put them into a column-vector labelled:
       mfs, so as to pick the relevant binary indicators, when needed. Cf. V.5, re below */
          READ DATA I_MF INTO mfs=[I_MF];
    /* II.2.2. Importing the binary indicators of the modes of the fixed factors characterizing strata
       and placing them into a matrix labelled: M_Mfs_id, strata (rows) X modes (columns)
       cf. V.2, here below*/
          READ DATA mf_id INTO Strata=[Stratum] {m in mfs} <M_mfs_id[Stratum,m]=col(m)>;
    /* II.2.3. Importing the values of the strata-specific predicted probabilities of in-time delivery
       cf. V.3., here below */
          READ DATA KPI_Co2Cd INTO Strata=[stratum] pi_s_t = SPD_by_T;
    /* II.2.4. Importing of the real mail weights of factors' modes
       Cf. V.4, here below */
          READ DATA RMW_Co2Cd INTO mfs=[I_MF] Mf_W=Omega_mf;
    /* II.3. Declaration of the unknown variables: i.e., the weights to be allocated to the strata of items
       sent from each Co to each Cd. Setting of their lower limit: all must be non-negative, i.e., >=0.
       N.B. Their upper limit (i.e., <=1) is naturally satisfied by the sum-constraint, defined in II.4 */
          VAR w{1..&N_strata} >=0; /* w_{s|Co2Cd} → w */
     /* II.4. Specification of the sum-constraint: strata weights must add up to unity */
          CONSTRAINT Overall_consistency: SUM{s in Strata} w[s]=1; /* Σ_{s∈S(Co2Cd)} w_{s|Co2Cd} = 1 */
    /* II.5. Specification of the constraints limiting the search to weighting schemes perfectly
       consistent with the conditional real mail modes' weights:
       Σ_{s∈S(Co2Cd)} w_{s|Co2Cd} × I_{m(f|C)}^{s|Co2Cd} = Ω_{m(f|C)} */
          CONSTRAINT RMW_Mf {m in mfs}:  ( SUM{s in strata} w[s]*M_mfs_id[s,m] ) = Mf_W[m];
```

```
        /* II.6. Definition of the OBJECTIVE: i.e., either MINimizing or MAXimizing the Co2Cd-KPI :
           Π̂(Co2Cd, 𝕥| 𝒘_Co2Cd) = ∑_{s∈S(Co2Cd)} 𝒘_{s|Co2Cd} × Π̂_{s,𝕥}  →  PI_Co2Cd  */
              &OBJECTIVE PI_Co2Cd = SUM{s in Strata} (pi_s_t[s]*w[s]);
           /* Specifies that the solution is to be found by the linear programming algorithm */
              SOLVE WITH LP;
        /* II.7. Saving the summary of the essential outcomes: the value of the OBJECTIVE at the optimum*/
              ODS OUTPUT ProblemSummary=Problem SolutionSummary=Optimum;
        /* II.8. EXPAND makes the constraints & objective function explicit, for checking purpose */
              EXPAND;
        /* II.9. Saving optimal weights: either  𝒘_Co2Cd, or  𝒘̄_Co2Cd  */
              CREATE DATA W_PI_s_t FROM [Stratum] w pi_s_t;
         QUIT;
%MEND Optimization;

/* III. Macro editing the results' files, called by the %Pooling-macro: cf. § IV */
%MACRO Edition;
      DATA Status; LENGTH Status_&OBJECTIVE $22.; SET Optimum;
            IF Label1="Solution Status" THEN Status_&OBJECTIVE=cValue1;
            IF Label1="Solution Status"; KEEP Status_&OBJECTIVE;
      RUN;
      DATA Opt_value; FORMAT Objective_&OBJECTIVE BEST14.; SET Optimum;
            IF Label1="Objective Value" THEN Objective_&OBJECTIVE=cValue1;
            IF Label1="Objective Value"; KEEP Objective_&OBJECTIVE;
      RUN;
      DATA Optimum_&OBJECTIVE; MERGE Status Opt_value; Co="&Co_path";Cd="&Cd_path"; RUN;
      DATA W_PI_s_t_&OBJECTIVE; SET W_PI_s_t; w_&OBJECTIVE=w;
            w_pi_s_t_&OBJECTIVE=w*pi_s_t; Co="&Co_path";Cd="&Cd_path"; DROP w;
      RUN;
%MEND Edition;

/* IV. Macro pooling Co2Cd optimal weighting schemes */
%MACRO Pooling;
        DATA Optima_Co2Cd; MERGE Optimum_MIN Optimum_MAX; BY Co Cd; RUN;
        DATA W_PI_s_t_Co2Cd; MERGE SWB_Co2Cd W_PI_s_t_MIN W_PI_s_t_MAX; RUN;
        %IF &p=1 %THEN %DO; %Edition;
                              DATA Optima_ALL; SET Optima_Co2Cd; RUN;
                              DATA W_PI_s_t_ALL; SET W_PI_s_t_Co2Cd; RUN;
                        %END;
        %ELSE %IF &p>1 %THEN %DO; %Edition;
                                    DATA Optima_ALL; SET Optima_ALL Optima_Co2Cd; RUN;
                                    DATA W_PI_s_t_ALL; SET W_PI_s_t_ALL W_PI_s_t_Co2Cd; RUN;
                              %END;
%MEND Pooling;
/* V. Macro governing the processing of the calculations of all vectors of weights
   minimizing/maximizing the point-estimates of the probabilities of in-time delivery, per Co2Cd-path */
%MACRO Looping_Co2Cd;
     %DO p = 1 %TO &N_paths;
       /* V.1. Selection of the strata standard weights specific to the  p^th  Co2Cd-path
          and ranking of these according to the factors determining the stratification */
          DATA SWB_Co2Cd; SET SWB; IF Path=&p; RUN;
          PROC SORT DATA=SWB_Co2Cd; BY Fk Pl Sw Ud Uo Wd; RUN;
          DATA SWB_Co2Cd; SET SWB_Co2Cd; Stratum=_N_; DROP Co2Cd Path; RUN;
       /* V.2. Generation of the binary indicators of whether, or not, each stratum is characterized by each of the
          factors' modes: matrix, to be imported in the OPTMODEL-procedure, cf. II.2.2, here above */
          PROC GLMMOD DATA=SWB_Co2Cd OUTDESIGN=mf_id PREFIX=I_MF NOPRINT;
                CLASS Fk Pl Sw Ud Uo Wd;
                MODEL Stratum = Fk Pl Sw Ud Uo Wd / NOINT;
          RUN;
```

```
        /* V.3. Selection of the strata-specific predicted probabilities of in-time delivery, for items sent from Co to Cd,
           to be imported in the OPTMODEL-procedure, cf. II.2.3., here above */
              DATA _NULL_; SET SWB_Co2Cd END=LAST;
                    IF LAST THEN DO; CALL SYMPUT("N_strata",LEFT(_N_));
                                          CALL SYMPUT("Co_path",Co); CALL SYMPUT("Cd_path",Cd);
                                    END;
              RUN;
              DATA KPI_Co2Cd; SET Performance_D5; IF (Co="&Co_path" AND Cd="&Cd_path"); RUN;
              PROC SORT DATA=KPI_Co2Cd; BY Fk Pl Sw Ud Uo Wd; RUN;
              DATA KPI_Co2Cd; SET KPI_Co2Cd; Stratum=_N_; KEEP Stratum SPD_by_T; RUN;
        /* V.4. Selection of the real mail weights of factors' modes relevant to the p^th Co2Cd-path examined,
           to be imported in the OPTMODEL-procedure, cf. II.2.4., here above */
              DATA RMW_Co2Cd; SET RMW;
                    IF (Country="&Co_path" AND Factor^="Ud") OR (Country="&Cd_path" AND Factor="Ud");
              RUN;
              PROC SORT DATA=RMW_Co2Cd; BY Factor Mode; RUN;
              DATA _NULL_; SET RMW_Co2Cd END=Last; IF Last THEN CALL SYMPUT("N_modes",LEFT(_N_)); RUN;
        /* V.5. Labelling factors' modes, consistent with that of their binary indicators (cf. § V.2, here above)
            to be imported in the OPTMODEL-procedure: cf. II.2.1., here above */
              DATA I_MF(DROP=m); LENGTH I_MF $6.;
                    DO m=1 TO &N_modes;
                       IF m<10 THEN I_MF = 'I_MF'||PUT(m,1.); ELSE I_MF = 'I_MF'||PUT(m,2.); OUTPUT;
                    END;
              RUN;
              DATA RMW_Co2Cd; MERGE I_MF RMW_Co2Cd; KEEP I_MF Omega_mf; RUN;
        /* V.6. Definition of the objectives and execution of the optimization */
              %LET OBJECTIVE = MAX; %Optimization; %Edition; %LET OBJECTIVE = MIN; %Optimization; %Edition;
        %Pooling;
    %END;
%MEND Looping_Co2Cd;
  /* Execution of the just described MACRO-routine */
% Looping_Co2Cd;
```

A.2 *P*rogram designed to fit the mixed ***MCL*** model and to test gaps between response thresholds: cf. *5.4.*

```
/* I. Input of test-items' records and categorization of delivery times into performance classes */
DATA EtE_records; LENGTH Performance $5.; INFILE 'C:\... ... \EtE_records.csv' DLM="," DSD;
      /* Response: Delivery_time.   Predictors: labels defined in Table 1 (fixed factors: f ∈ ℱ).
         Zo and Zd denote logistic areas of origin/destination, transformed into dummies:
         Z^O_iα and Z^D_iδ in (5), through the CLASS-statement in the PROC GLIMMIX (cf. here below: II.2) */
      INPUT Delivery_time Co $ Cd $ Uo $ Ud $ Wd $ Sw $ Fk $ Pl $ Zo $ Zd $;
      /* Categorization into ordered performance classes */
      IF Delivery_time=1 THEN Performance = "T_1"; /* 𝕥 = 1 */
         ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
            ELSE IF Delivery_time=9 THEN Performance = "T_9"; /* 𝕥 = 9 */
               ELSE IF Delivery_time=10 THEN Performance = "T_X"; /* 𝕥 = 10 */
                  ELSE IF Delivery_time>10 THEN Performance = "T_X+"; /* 𝕥 = 𝕥̄ + 1 ≡ 𝕃 */
RUN;
/* II. Mixed Multinomial Cumulative Logistic (MCL) regression */
     /* II.1. Definition of nominal factors to be entered as predictors in the model and choice of their base-levels
           i.e., arbitrarily here: AT, Austria, for both countries of origin and destination,
              and Rt, rural towns, for degrees of urbanization of outbound and inbound areas */
        /* Geographical features */
           %LET Geo = Co  Cd  Uo  Ud;
           %LET Geo_Ref = Co(REF='AT')  Cd(REF='AT')  Uo(REF='Rt')  Ud(REF='Rt');
        /* Mail characteristics: Ad excluded because the addressing mode has become non discriminant */
           %LET MCs = Wd  Sw  Fk  Pl;
           %LET MCs_ref= Wd(REF='6_Sa')  Sw(REF='C6_20g')  Fk(REF='St')  Pl(REF='Pu');
```

```
/* II.2. Procedure fitting Generalized Linear Mixed Models (GLIMMIX) */
   /* Estimation option: maximization of the likelihood using Laplace approximation method */
     PROC GLIMMIX DATA=EtE_records METHOD=LAPLACE ITDETAILS GRADIENT;
       /* Declaration of fixed predictors and random components as classification indicators */
       /* Automatic coding into binary dummies: x^f_{i,m(f)}, Z^O_{i,o} and Z^D_{i,d} */
       CLASS  &Geo_Ref  &MCs_ref  Zo  Zd;
      /* Model specification, according to equation (5),
        - with the COVB-option to save the covariance matrix of the fixed-effects parameter estimates */
       MODEL Performance = &Geo &MCs /
                         LINK=Cumlogit DIST=Multinomial DDFM=BW SOLUTION COVB;
      - including the random intercepts, encompassing the spatial heterogeneity in logistics */
       RANDOM INTERCEPT / SUBJECT=Zo TYPE=VC; RANDOM INTERCEPT / SUBJECT=Zd TYPE=VC;
     /* Combination of non-linear optimization options to ensure convergence
        within a reasonable time limit */
       NLOPTIONS MAXITER=10000 MAXFUNC=10000 ABSFTOL=0.00001
                  GCONV=0 ABSGCONV = 1e-18 FCONV=0 TECHNIQUE=NRRIDG;
    /* CONTRAST-statements customizing hypothesis tests,
      here designed to compare the standardized θ_{t|B}-thresholds defined by cf. (15.d) */
      CONTRAST ' T_1 vs T_2'  Intercept  1  -1  0  0  0  0  0  0  0  0;
      CONTRAST ' T_2 vs T_3'  Intercept  0   1  -1  0  0  0  0  0  0  0;
      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
      CONTRAST ' T_9  vs T_X' Intercept  0   0   0  0  0  0  0  0  1  -1;
    /* Filing of standardized parameters' estimates differentiating fixed factors' modes:
      in (15.e) and of results of tests of global fixed effects */
      ODS OUTPUT PARAMETERESTIMATES=Estimates TESTS3=Fix_eff;
    /* Filing of (a) the estimated standardized variances in effects of random components in:
      V_random_effects; (b) the matrix of estimated variances of, and covariances between, estimates of
      the other parameters in: VCov; (c) and the output from tests of differences between the
      standardized thresholds in: Diff_intercepts. */
      ODS OUTPUT COVPARMS=V_random_effects COVB=VCov CONTRASTS=Diff_intercepts;
    RUN;
  /* Excerpting estimates of standardized θ_{t|B}-thresholds from the file of parameters' estimates */
    DATA EST_Thresholds;SET Estimates;IF Effect="Intercept";KEEP Performance Estimate; RUN;
  /* Excerpting estimated variances of, and covariances between, estimates of standardized thresholds,
    from the entire VCov-matrix of parameters' estimates */
    DATA VCov_Thresholds; SET VCov; IF Effect="Intercept"; DROP Effect &Geo &MCs Row; RUN;

/* III. Routine calculating all statistics relevant to contrast estimates of successive θ_{t|B}-thresholds,
      not provided by the execution of the CONTRASTS-statements: cf. § 5.4.2, Table 8 */
  %LET UT=10; /* Upper limit on delivery time: t̄ */
  /* Programming of the tests of differences between the θ_{t|B}-thresholds' estimates */
  %MACRO Cut_points;
      PROC IML; /* Use of the Interactive Matrix Language (IML) */
        /* Logging of thresholds' estimates into a column-vector, named: V_est_thresh */
        USE EST_Thresholds; READ ALL VAR _NUM_ INTO V_est_thresh;
        Theta_est = − V_est_thresh[1:&UT];
        /* Logging of compared thresholds' names into column-vectors: Theta_t and Theta_tp1 */
        READ ALL VAR _CHAR_ INTO Parms;   Theta_t=Parms[1:&UT −1]; Theta_tp1=Parms[2:&UT];
        /* Naming of pairs of thresholds' estimates */
        Theta_compared=Theta_t||Theta_tp1;
        THETAs = {"THETA_t","THETA_tp1"};
        CREATE Pairs FROM Theta_compared [COLNAME=THETAs]; APPEND FROM Theta_compared;
        USE VCov_Thresholds;
        /* Logging of the estimated variances of, and covariances between, the estimates of
          the standardized thresholds into a matrix, named: M_vcov */
        READ ALL VAR _NUM_ INTO M_vcov; VCOV = M_vcov[1:&UT,1:&UT];
```

```
                    /* Dimensioning and initialization of intermediate and final outputs vectors and matrix*/
                    Stderror=SHAPE(0, &UT−1,1); Dif=SHAPE(0,&UT−1,1); M_test=SHAPE(0,&UT−1,4);
                    /* Comparisons between thresholds' estimates */
                    DO t=1 TO &UT−1;
                        /* Differences between thresholds' estimates */
                        Dif[t]=Theta_est[t]−Theta_est[t+1];
                        /* Standard errors of differences between thresholds' estimates */
                        Stderror[t]=(VCOV[t,t]+VCOV[t+1,t+1]−VCOV[t,t+1]−VCOV[t+1,t])**0.5;
                    END;
                    /*  z-statistic and  p-values assessing the significance of the differences */
                    z_stat=Dif/Stderror; p_value=1−PROBNORM(ABS(z_stat));
                    /* Logging results of the tests of differences */
                    M_test[,1]=Dif;M_test[,2]=Stderror; M_test[,3]=z_stat; M_test[,4]=p_value;
                    Labels={"Difference","StdError","z-Stat","p-Value"};
                    CREATE Test_file FROM M_test [COLNAME=Labels]; APPEND FROM M_test; CLOSE Test_file;
                QUIT;
                DATA Thresholds_tests; MERGE Pairs Test_file; RUN;
    %MEND Cut_points;
    /* Execution of the just described MACRO-routine */
 %Cut_points;
```

A.3 Program designed to fit the ***NB*** model and therefrom estimate in-time delivery probabilities: cf. § 4.1.1.

```
/* I. Negative Binomial (NB) regression: statements similar to those programmed in: II.2. of annex A.2 */
PROC GLIMMIX DATA=EtE_records METHOD=LAPLACE ITDETAILS GRADIENT;
        CLASS &Geo_Ref   &MCs_ref   Zo Zd;
        /* Response:  N_i = T_i − 1, labelled: N_count */
        MODEL N_count = &Countries &Zones &MCs / SOLUTION DDFM=BW DIST=NEGBIN LINK=LOG;
        NLOPTIONS MAXITER=10000 MAXFUNC=10000 ... TECHNIQUE=NRRIDG; /* cf. II.2. */
        /* Filing of the Marginal Linear Predictor: MLP= ln μ_i = −Q̂_i, cf. formula: (11).
           The NOBLUP-option neutralizes the effects of random components */
        OUTPUT OUT= NB_Predictions PRED(NOBLUP)=MLP;
        /* Filing of:
           (a) parameters' estimates,
           (b) the outputs from the tests of the global effects of fixed factors:
                F-statistics in Table 7  →  Tests3= Global_fixed_effects
           (c) the CovParms-output includes the estimated variances of the random components
               and of the overdispersion-parameter specific to the NB-distribution */
        ODS OUTPUT ParameterEstimates=Param_est Tests3=Global_fixed_effects CovParms=CovParms;
RUN;
/* Saving φ̂ (Phi) - called scale of the NB distribution - as a SAS macro variable */
DATA _NULL_; SET CovParms; IF CovParm = "Scale" THEN CALL SYMPUT("Phi", Estimate); RUN;

/* II. Routine inferring strata-level predicted probabilities of
   (a) on-time delivery:  π̂_{s,t}, labelled: PdinT(t) ; (b) and in-time (i.e., within deadline):  Π̂_{s,t}, labelled: PDbyT(t)   */
%MACRO NB_Cpmf;
            PROC SORT DATA=NB_Predictions NODUPKEYS;   BY &Geo &MCs;   RUN;
            %LET Phi_inverse = %SYSEVALF(1/&Phi);
            DATA NBD_Probabilities; SET NB_Predictions;
                    ARRAY PDinT{10} PDinT_1-PDinT_10; ARRAY PDbyT{10} PDbyT_1-PDbyT_10;
                    Q  = − MLP; Mu = EXP(− Q); p = 1/(1+ &Phi*Mu);
                    DO t = 1 TO 10; t_1 = t−1;
                        PDinT[t] = PDF('NEGBINOMIAL',t_1,p,&Phi_inverse);   /*  π̂_{s,t}  */
                        PDbyT[t] = CDF('NEGBINOMIAL',t_1,p,&Phi_inverse);   /*  Π̂_{s,t}  */
                    END;
            RUN;
%MEND NB_Cpmf;
/* Execution of the just described MACRO-routine */ %NB_Cpmf;
```

A.4 Program designed to derive the distributions of the estimates of the aggregate KPIs: cf. *6.1*.

```
/* I. Key constants dimensioning the simulation run */
/* I.1. Size of the parent sample: n */
    %LET Sample_size=105889;
/* I.2. Setting the number of bootstrap-samples to be generated by random resampling of the parent sampling */
    %LET NBS=10000;
/* II. Filing of postal outbound/inbound areas: from screening the data, which include Zo and Zd */
    Z^O_.σ ⟶ in: OUT_clusters ;  Z^D_i.d ⟶ in: IN_clusters */
    PROC SORT NODUPKEYS DATA=EtE_records OUT=OUT_clusters;  BY Zo;  RUN;
    DATA OUT_clusters;  SET OUT_clusters;  KEEP Zo;  RUN;
    PROC SORT NODUPKEYS DATA =EtE_records OUT= IN_clusters;  BY Zd;  RUN;
    DATA IN_clusters; SET IN_clusters;  KEEP Zd;  RUN;
/* III. MACRO routine generating local random variations in outbound/outbound logistics: [υ̂^O_σ/σ]^(b), [υ̂^D_d/σ]^(b) */
    %MACRO LRV;
        /* Outbound, labelled: upsilon_o; standard-deviation estimate recovered in: IV.3, below */
        DATA OUT_Clusters; SET OUT_Clusters;
            upsilon_o = RAND('NORMAL', 0 , &STDV_OUT);  /*  [υ̂^O_σ/σ]^(b) ~ 𝒩(0, ς̂^O/σ^(b))  */
        RUN;
        /* Inbound, labelled: upsilon_d; standard-deviation estimate recovered in: IV.3, below */
        DATA IN_Clusters; SET IN_Clusters;
            upsilon_d=RAND('NORMAL', 0 , &STDV_IN);  /*  [υ̂^D_d/σ]^(b) ~ 𝒩(0, ς̂^D/σ^(b))  */
        RUN;
    %MEND LRV;
/* IV. MACRO routine fitting the Multinomial Cumulative Logit model to a bootstrapped sample,
    resulting from a random selection of data in the parent sample: cf. VII, here after */
    %MACRO MCL_parametrization;
      /* IV.1. Estimation procedure adapted from II.2. in annex A.2 */
        PROC GLIMMIX DATA=Sample METHOD=LAPLACE ITDETAILS GRADIENT;
            CLASS &Geo_Ref &MCs_ref Zo Zd;
            MODEL Performance = &Geo &MCs /
                                DDFM=BW LINK=CUMLOGIT DIST=MULTINOMIAL SOLUTION;
            RANDOM INTERCEPT / SUBJECT=Zo TYPE=VC;  RANDOM INTERCEPT / SUBJECT=Zd TYPE=VC;
            NLOPTIONS MAXITER=10000 MAXFUNC=10000 ABSFTOL=0.00001
                      GCONV=0 ABSGCONV=1e-18 FCONV=0 TECHNIQUE=NRRIDG;
            ODS OUTPUT IterHistory=Iter_RAND ParameterEstimates=Estimates_RAND Tests3=Fix_eff_RAND;
            ODS OUTPUT CovParms=V_random_effects ConvergenceStatus=Convergence;
          /* Storing estimates whose values can next be recovered to produce predictions: cf. VI.1, here after */
            STORE SASUSER.MCL;
        RUN;
      /* IV.2. Message issued in case of divergence of the GLIMMIX-procedure */
        DATA _NULL_; SET Convergence; CALL SYMPUT('Divergence',Status);  RUN;
       /* Issuing, in the debugging log, of a message about the conver-/diver-gence of the GLIMMIX-procedure */
        %IF &Divergence=1 %THEN %PUT Divergence; %ELSE %PUT Convergence;
      /* IV.3. Recovery of the estimates of random components' standard deviations */
        DATA _NULL_; SET V_random_effects;
            IF Subject="Zo" THEN CALL SYMPUT('V_Outbound',Estimate);
            ELSE IF Subject="Zd" THEN CALL SYMPUT('V_Inbound',Estimate);
        RUN;
        %LET STDV_OUT=%SYSEVALF(&V_Outbound**0.5); %LET STDV_IN=%SYSEVALF(&V_inbound**0.5);
      /* IV.4. Call to the LRV-MACRO, commented in III, here above */ %LRV;
%MEND MCL_parametrization;
```

```
/* V. MACRO routine called by the Predictions-MACRO to finalize the estimation of the aggregate KPIs:
  Π̲̂(Eu_stream, 𝕥), Π̂(Eu_stream, 𝕥), Π̂̄(Eu_stream, 𝕥), for Eu_stream ∈ {Co2Eu, Eu2Cd, Eu2Eu}   */
%MACRO Poststratification;
        /* Ex-post weighting of the various sets of estimated probabilities of intime delivery: PD_by_T,
           considering, versus not, the local random variations: LRV_ versus NOLRV_-prefixes */
        DATA LPF; SET LPF;
              /* Using weights maximizing the Π̂(Co2Cd, 𝕥| 𝒲_Co2Cd)   */
              NOLRV_W_MIN_PDT_&Eu_stream=NOLRV_PD_by_T*W_MIN_&Eu_stream;
              LRV_W_MIN_PDT_&Eu_stream=LRV_PD_by_T*W_MIN_&Eu_stream;
              /* Using the standard weighting basis (SWB)   */
              NOLRV_W_SWB_PDT_&Eu_stream=NOLRV_PD_by_T*W_SWB_&Eu_stream;
              LRV_W_SWB_PDT_&Eu_stream=LRV_PD_by_T*W_SWB_&Eu_stream;
              /* Using weights maximizing the Π̂(Co2Cd, 𝕥| 𝒲_Co2Cd)   */
              NOLRV_W_MAX_PDT_&Eu_stream=NOLRV_PD_by_T*W_MAX_&Eu_stream;
              LRV_W_MAX_PDT_&Eu_stream=LRV_PD_by_T*W_MAX_&Eu_stream;
        RUN;
        /* Calculations of KPIs by summation of weighted estimates of the probabilities of intime delivery */
        PROC SORT DATA=LPF; BY &SortBy; RUN;
        PROC MEANS DATA=LPF NOPRINT; BY &SortBy;
              VAR NOLRV_W_MIN_PDT_&Eu_stream NOLRV_W_SWB_PDT_&Eu_stream
                  NOLRV_W_MAX_PDT_&Eu_stream LRV_W_MIN_PDT_&Eu_stream
                  LRV_W_SWB_PDT_&Eu_stream LRV_W_MAX_PDT_&Eu_stream;
              OUTPUT OUT=KPI_&Eu_stream
                        SUM = NOLRV_MIN_KPI_&Eu_stream NOLRV_SWB_KPI_&Eu_stream
                              NOLRV_MAX_KPI_&Eu_stream LRV_MIN_KPI_&Eu_stream
                              LRV_SWB_KPI_&Eu_stream LRV_MAX_KPI_&Eu_stream;
        RUN;
%MEND Poststratification;

/* VI. MACRO routine designed to predict the aggregate probabilities of in-time delivery */
%MACRO Predictions;
     /* VI.1. Extrapolation of the latent QoS-values specific to the various strata */
       The Postfitting Linear Model (PLM) procedure, used here after,
       (a) first, recovers parameters' estimates through its RESTORE option, specifying where they were stored,
           via a STORE-statement included in a previous model fitting procedure: cf. IV.1, here above;
       (b) next, it uses them to extrapolate responses corresponding to data on predictors,
           through a SCORE-statement which specifies the source (DATA-option) of these data (previously
           stored), as well as the output file (OUT-option).
        PROC PLM RESTORE=SASUSER.MCL NOINFO;
        /* The DATA- keyword in the SCORE statement refers to the file named Universe, describing the
           profiles of all the strata defining the postal universe, including their weights: ω_s|Co2Cd [Table 2],
           w̲_s|Co2Cd and w̄_s|Co2Cd [ Table 4], as well as those derived from these sets, relevant to aggregate
           KPIs obtained using the 𝕧_o2d   */
             SCORE DATA=Universe OUT=LPF PREDICTED=Score;
        /* The PREDICTED keyword in the SCORE statement requires extrapolations of the so-called linear
           predictor function (l.p.f.) values: i.e., the latent construct q̂_s|𝕥 defined by (15.3)-(15.5).*/
       RUN;
     /* VI.2. Insertion - in the LPF-file - of the simulated local random disturbances (cf. III, here above) */
       PROC SORT DATA=LPF; BY Zo; RUN; DATA LPF; MERGE LPF OUT_Clusters; BY Zo; RUN;
       PROC SORT DATA=LPF; BY Zd; RUN; DATA LPF; MERGE LPF IN_Clusters; BY Zd; RUN;
     /* VI.3. Calculations of the estimates of the strata-level probabilities of in-time delivery: Π̂_s,𝕥,
              by application of formula (17), cf. § 4.2.2. */
       DATA LPF; SET LPF; Performance_level=_LEVEL_;
              /* Taking local random variations in logistics into account */
              LRV_PD_by_T = 1/(1+EXP(−(Score+upsilon_o+upsilon_d)));
              /* Ignoring local random variations in logistics */ NOLRV_PD_by_T = 1/(1+EXP(−Score));
       RUN;
```

```
        /* VI.4. Computations of the estimates of aggregate KPIs,
           through execution of the Poststratification-MACRO: cf. V, here above */
        %LET Eu_stream = Eu2Eu;
                %LET SortBy = Performance_level;
                %Poststratification;
        %LET Eu_stream = Co2Eu;
                %LET SortBy = Co Performance_level;
                %Poststratification;
        %LET Eu_stream = Eu2Cd;
                %LET SortBy = Cd Performance_level;
                % Poststratification;
%MEND Predictions;
/* VII. MACRO governing the resampling and replicating the estimation of the aggregate KPIs
   on each  b  bootstrap-sample */
%MACRO Simulation;
            %DO b=1 %TO &N_samples;
            /* Unrestricted Random Sampling (option abbreviated: URS) - i.e., with replacement - of real test-records,
               selected from the parent sample of size: n, by the SURVEYSELECT- procedure */
                PROC SURVEYSELECT
                        DATA=EtE_records METHOD=URS n=&Sample_size OUTHITS OUT=Sample NOPRINT;
                RUN;
            /* Call to the routine fitting the MCL to the bootstrap sample: cf. IV, here above */
                %MCL_parametrization;
             /* Call to the routine estimating aggregate KPIs: cf. VI, here above */
                %Predictions;
            /* Stockpiling of KPIs' estimates */
                %IF &b=1 %THEN %DO;
                                        DATA BOOTSTRAP_Eu2Eu; SET KPI_Eu2Eu; RUN;
                                        DATA BOOTSTRAP_Co2Eu; SET KPI_Co2Eu; RUN;
                                        DATA BOOTSTRAP_Eu2Cd; SET KPI_Eu2Cd; RUN;
                                %END;
                %ELSE %DO;
                                DATA BOOTSTRAP_Eu2Eu; SET BOOTSTRAP_Eu2Eu KPI_Eu2Eu; RUN;
                                DATA BOOTSTRAP_Co2Eu; SET BOOTSTRAP_Co2Eu KPI_Co2Eu; RUN;
                                DATA BOOTSTRAP_Eu2Cd; SET BOOTSTRAP_Eu2Cd KPI_Eu2Cd; RUN;
                        %END;
            %END;
%MEND Simulation;

/* VIII. Execution statement running the bootstrapping MACRO */
%Simulation;
```

References

Abdel-Aty, M. A., and Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention* 32(5), 633-642. https://doi.org/10.1016/S0001-4575(99)00094-9

Abts, K. C., Ivy, J. A., and DeWoody, J. A. (2018). Demographic, environmental and genetic determinants of mating success in captive koalas (Phascolarctos cinereus). *Zoo Biology* 37(6), 416–433. https://doi.org/10.1002/zoo.21457

Agresti, A. (2002). *Categorical Data Analysis*, John Wiley & Sons, Inc. DOI:10.1002/0471249688

Agresti, A. (2007, 2019). *An Introduction to Categorical Data Analysis*, John Wiley & Sons, Inc.Alam, N., and Lee Ng, S. (2014). Banking mergers - an application of matching strategy. *Review of Accounting and Finance* 13(1), 2-23. https://doi.org/10.1108/RAF-12-2012-0124

Amrhein, V., Greenland, S., and McShane, B. (2019). Retire statistical significance. *Nature* 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9

Anderson, J. A., and Philips, P. R. (1981). Regression, discrimination and measurement models of ordered categorical variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 30(1), 22–31. https://doi.org/10.2307/2346654

Armstrong, B. G., and Sloan, M. (1989). Ordinal Regression Models for Epidemiologic Data. *American Journal of Epidemiology* 129(1), 191–204. https://doi.org/10.1093/oxfordjournals.aje.a115109

Anderson, J. A., and Philips, P. R. (1981). Regression, discrimination and measurement models of ordered categorical variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 30(1), 22–31. https://doi.org/10.2307/2346654

Austin, P. C., and Leckie, L. (2020) Bootstrapped inference for variance parameters, measures of heterogeneity and random effects in multilevel logistic regression models, *Journal of Statistical Computation and Simulation*, 90(17), 3175-3199. https://doi.org/10.1080/00949655.2020.1797738

Barmby, T., Nolan, M., and Winkelmann, R. (2001). Contracted workdays and absence. *The Manchester School* 69(3), 269-275. https://doi.org/10.1111/1467-9957.00247

Bermúdez, L., Karlis, D., and Santolino, M. (2018). A discrete mixture regression for modeling the duration of nonhospitalization medical leave of motor accident victims. *Accident Analysis and Prevention* 121, 157–165. https://doi.org/10.1016/j.aap.2018.09.006

Betensky, R. (2019). The p-value requires context, not a threshold. *The American Statistician* 73(1), 115–117. ttps://doi.org/10.1080/00031305.2018.1529624

Boysen, N., Fedtke, S., and Schwerdfeger, S. (2021). Last-mile delivery concepts: a survey from an operational research perspective. *OR Spectrum* 43, 1–58. https://doi.org/10.1007/s00291-020-00607-8

Brusco, M. (2022). Logistic Regression via Excel Spreadsheets: Mechanics, Model Selection, and Relative Predictor Importance. *INFORMS Transactions on Education* 23(1), 1-11. https://pubsonline.informs.org/doi/abs/10.1287/ited.2021.0263

Bultez, A., and Naert, P. (1979). Does Lag Structure Really Matter in Optimizing Advertising Expenditures? *Management Science*, 25(5). 454-465. https://doi.org/10.1287/mnsc.32.2.182

Bultez, A., Derbaix, C., and Herrmann, J.-L. (2022). Statistically significant? Let us recognize that estimates of tested effects are uncertain. *Recherche et Applications En Marketing (English Edition)*, 37(1), 82-105. https://doi.org/10.1177/20515707211040743

Bultez, A., and Herrmann, J.-L. (2025). Value added to marketing research diagnoses by add-ons to p-values. *Journal of Marketing Analytics*, forthcoming. https://doi.org/10.1057/s41270-024-00351-w

Bultez, A., Laurent, G., and Lemay, L. (2025). Quantifying relationships between ordinal categorical variables. Application to metrics tracked by satisfaction barometers. *Recherche et Applications En Marketing (English Edition)*, forthcoming. https://www.researchgate.net/publication/388217974_...

Calabrese, R., Marra, G., and Osmetti, S. A. (2016). Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. *The Journal of the Operational Research Society*, 67(4), 604-615. https://doi.org/10.1057/jors.2015.64

Calin-Jageman, R. J., and Cumming, G. (2019). The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known. *The American Statistician* 73(sup 1), 271-280. https://doi.org/10.1080/00031305.2018.1518266

Cai, B., and Shimizu, I. (2014). Negative Binomials Regression Model in Analysis of Wait Time at Hospital Emergency Department. *Proceedings of the American Statistical Association* (April), 0:4262. PMCID: PMC7183738 https://pubmed.ncbi.nlm.nih.gov/32336961/

Carrubba, C., Friedman, B., Martin, A. D., and Vanberg, G. (2012). Who Controls the Content of Supreme Court Opinions? *American Journal of Political Science* 56(2), 400–441. https://doi.org/10.1111/j.1540-5907.2011.00557

Castillo, V. E., Mollenkopf, D. A., Bell, J. E., and Bozdogan, H. (2018). Supply Chain Integrity: A Key to Sustainable Supply Chain Management. *Journal of Business Logistics* 39(1), 38–56. https://doi.org/10.1111/jbl.12176

Caulkins, J.P., Barnett, A., Larkey, P.D., Yuan, Y., and Goranson, J. (1993). The On-Time Machines: Some Analyses of Airline Punctuality. *Operations Research*, 41(4), 710-720. https://doi.org/10.1287/opre.41.4.710

CEN: EUROPEAN COMMITTEE FOR STANDARDIZATION. Technical Committee CEN/TC 331, EN 13850 (2020), Postal services - Quality of services -Measurement of the transit time of end-to-end services for single piece priority mail and first class mail. https://www.cencenelec.eu/about-cen/

Dalla V. L., Leisen, F., Rossini, L., and Zhu, W. (2020). Bayesian analysis of immigration in Europe with generalized logistic regression. *Journal of Applied Statistics* 47(3), 424-438. https://doi.org/10.1080/02664763.2019.1642310

De Haan, E., Verhoef, P., and Wiesel, T. (2015). The Predictive Ability of Different Customer Feedback Metrics for Retention. *International Journal of Research in Marketing* 32(2), 195-206. https://doi.org/10.1016/j.ijresmar.2015.02.004

Denuit, M., Hainaut, D., and Trufin, J. (2019). *Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions*, Springer Nature Switzerland AG. https://link.springer.com/book/10.1007/978-3-030-25820-7

Diekmann, A., Bruderer, E. H., Hartmann, J., Kurz, K., Liebe, U., and Preisendörfer, P. (2022), Environmental Inequality in Four European Cities: A Study Combining Household Survey and Geo-Referenced Data. *European Sociological Review* 39(1), 44-66. https://doi.org/10.1093/esr/jcac028

Dinler, N., and Rankin, W. B. (2020). Increasing Airports' On-Time Arrival Performance Through Airport Capacity and Efficiency Indicators. *International Journal of Aviation, Aeronautics, and Aerospace* 7(1). https://doi.org/10.15394/ijaaa.2020.1444

Efron, B., and Tibshirani, R. (1994). *An Introduction to Bootstrap*. Chapman & Hall /CRC (New York). https://doi.org/10.1201/9780429246593

Efron, B., and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science* 1(1), 54-75. https://doi.org/10.1214/ss/1177013815

Ehrenberg, A. S. C. (1959). The pattern of consumer purchases. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 8(1), 26-41. https://doi.org/10.2307/2985810

Farid, A., and Ksaibati, K. (2021). Modeling severities of motorcycle crashes using random parameters. *Journal of Traffic and Transportation Engineering* 8(2), 225-236. https://doi.org/10.1016/j.jtte.2020.01.001

Guadagni, P. M., and Little, J. D. C. (1983). A Logit Model of Brand Choice Calibrated on Scanner Data. *Marketing Science* 2(3), 203-238. https://doi.org/10.1287/mksc.2.3.203

Harrell, F. E. Jr. (2015), *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer International Publishing (Switzerland). https://link.springer.com/book/10.1007/978-3-319-19425-7

Hilbe, J. M. (2011). *Negative Binomial Regression*, Cambridge University Press. https://doi.org/10.1017/CBO9780511973420

Hurlbert, S.H., Levine, R.A., and Utts. J. (2019). Coup de Grâce for a tough old bull: 'Statistically significant' expires. *The American Statistician* 73: 352–357. https://doi.org/10.1080/00031305.2018.1543616

Hoang, V.-N., and Watson, J. (2022), Teaching binary logistic regression modeling in an introductory business analytics course. *Decision Sciences Journal of Innovative Education* 20(4), 201–211. https://doi.org/10.1111/dsji.12274

IPC (March 2024). INTERNATIONAL MAIL QUALITY OF SERVICE MONITORING: UNEX™ CEN 2023, 1-10. https://www.ipc.be/services/operational-performance-services/unex/results

Khan, R. J., Gebreab, S. Y., Sims, M., Riestra, P., Xu1, R., and Sharon K Davis, S. K. (2015). Prevalence, associated factors and heritabilities of metabolic syndrome and its individual components in African Americans: the Jackson Heart Study. *British Medical Journal Open*, 5(10). https://bmjopen.bmj.com/content/5/10/e008675

Kim, Y., Choi, Y. K., and Emery, S. (2013). Logistic Regression With Multiple Random Effects: A Simulation Study of Estimation Methods and Statistical Packages. *The American Statistician* 67(3), 171-182. http://www.jstor.org/stable/24591462

Kiernan, K., Tao, J., and Gibbs, P. (2012). Tips and Strategies for Mixed Modeling with SAS/STAT® Procedures. Paper 332-2012, SAS Global Forum Proceedings, SAS Institute Inc., Cary, NC, USA. https://support.sas.com/resources/papers/proceedings12/332-2012.pdf

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*, Series B (Methodological), 42(2), 109–142. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

Mood, C. (2010). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About it. *European Sociological Review* 26(1), 67–82. https://doi.org/10.1093/esr/jcp006

Mulder, J., and Wagenmakers, E-J. (eds) (2016). Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments. *The Journal of Mathematical Psychology* 72(June), 1–220. https://psycnet.apa.org/doi/10.1016/j.jmp.2016.01.002

Norton, E. C., Dowd, B. E., and Matthew L. Maciejew, M. L. (2018). Odds Ratios - Current Best Practice and Use, *Journal of the American Medical Association* 320(1), 84-85. https://pubmed.ncbi.nlm.nih.gov/29971384/

Papadopoulos, A., Roland, B., and Stark, R. B. (2021). Does Home Health Care Increase the Probability of 30-Day Hospital Readmissions? Interpreting Coefficient Sign Reversals, or Their Absence, in Binary Logistic Regression Analysis. *The American Statistician*, 75(2), 173-184. https://doi.org/10.1080/00031305.2019.1704873

Peng, C.-Y. J., So, T.-S. H., Stage, F. K., Edward P. St., and John, E. P. (2002). The Use and Interpretation of Logistic Regression in Higher Education Journals: 1988-1999. *Research in Higher Education*, 43(3), 259-293. https://doi.org/10.1023/A:1014858517172

Pritchard, J. P., Slovic, A. D., Giannotti, M., Geurs, K., Nardocci, A., Hagen-Zanker, A., Diego, B., Tomasiello, D. B., and Kumar, P. (2021). Satisfaction with travel, ideal commuting, and accessibility to employment. *Journal of Transport and Land Use* 14(1), 995-1017. http://dx.doi.org/10.5198/jtlu.2021.1835

Rao, J. N. K., Molina, I. (2015). *Small Area Estimation,* John Wiley & Sons, ISBN: 978-1-118-73578-7

Ross, M. L., Lawston, A. N., Lowsky, L. O., and Hackman, C. L. (2022). What Factors Predict COVID-19 Vaccine Uptake Intention in College Students? *American Journal of Health Education*, *53*(4), 237–247. https://doi.org/10.1080/19325037.2022

Shaban, T. F., and Alkawareek, M. Y. (2022). Prediction of qualitative antibiofilm activity of antibiotics using supervised machine learning techniques. *Computers in Biology and Medicine* 140, 1-9, available online. https://doi.org/10.1016/j.compbiomed.2021.105065

Simatupang, T. M., Wright, A. C., and Sridharan, R. (2002). The knowledge of coordination for supply chain integration. *Business Process Management Journal*, 8(3) 289-308. https://doi.org/10.1108/14637150210428989

Solow, R. M. (1960). On a Family of Lag Distributions. *Econometrica* 28(2), 393-406. https://doi.org/0012-9682(196004)28:2%3C393:OAFOLD%3E2.0.CO;2-0

Stevens, S. S. (1946). On the theory of scales of measurement. *Science* 103 (2684), 677-680. https://www.science.org/doi/10.1126/science.103.2684.677

Stoklosa, J., Blakey, R. V., and Hui, F. C. K. (2022) An Overview of Modern Applications of Negative Binomial Modelling in Ecology and Biodiversity Diversity. 14(320), 1-25. https://doi.org/10.3390/d14050320

Sugasawa, S., and Kubokawa, T. (2023). *Mixed-Effects Models and Small Area Estimation*, Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-19-9486-9

UPU: Universal Postal Union. (2023). State of the Postal Sector 2023. A Hyper-Collaborative Path to Postal Development, Berne (CH). https://www.upu.int/UPU/media/upu/publications/State-of-the-Postal-Sector-2023.pdf

Uusitalo, J., Ylhäisi, O., Rummukainen, H., and Makkonen, M. (2018) Predicting probability of A-quality lumber of Scots pine (Pinus sylvestris L.) prior to or concurrently with logging operation, *Scandinavian Journal of Forest Research*, 33(5), 475-483. https://doi.org/10.1080/02827581.2018.1461922

Wang, K.-S., Owusu, D., Yue, P. Y., and Xie, C. (2016). Bayesian logistic regression in detection of gene–steroid interaction for cancer at PDLIM5 locus. *Journal of Genetics* 95(2), 331-340. https://doi.org/10.1007/s12041-016-0642-1

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician* 70(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

White, R. E., Pearson, J. N., and Wilson, J. R. (1999). JIT Manufacturing: A Survey of Implementations in Small and Large U.S. Manufacturers. *Management Science*, 45(1),1-15.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Springer-Verlag (Berlin). https://link.springer.com/book/10.1007/978-3-540-78389-3

Wolter, K., Jergovic, D., Moore, W., Murphy, J., and O'Muircheartaigh, C. (2003). Reliability of the Uncertified Ballots in the 2000 Presidential Election in Florida. *The American Statistician*, 57(1), 1-14. https://doi.org/10.1198/0003130031144

Yang, Y., Hu, X., and Jiang, H. (2022). Group penalized logistic regressions predict up and down trends for stock prices. *North American Journal of Economics and Finance*, 59(January), 1-15. https://doi.org/10.1016/j.najef.2021.101564

Yirga, A. A., Melesse, S. F., Mwambi, H. G., and Ayele, D. G. (2020). Negative binomial mixed models for analyzing longitudinal CD4 count data. *Nature, Scientific Reports* 10(16742), open access. https ://doi.org/10.1038/s4159 8-020-73883 -7

Zellner, A. (2002). Keep it sophisticatedly simple. In: Zellner A, Keuzenkamp HA, McAleer M, eds. *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*, Cambridge University Press: 242-262. https://www.cambridge.org/core/books/simplicity-inference-and-modelling/A03DB5BC14FFA1662EBCE25316F460FE

Zhang, F., Sun, B., Diao, X., Zhao, W., and Shu, T. (2021). Prediction of adverse drug reactions based on knowledge graph embedding. *BMC Medical Informatics and Decision Making* 21(38), 1-11. https://doi.org/10.1186/s12911-021-01402-3