Towards a Reference Methodological Framework for processing MNO data for Official Statistics

Fabio Ricciato
EUROSTAT Task Force on Big Data

Abstract

Mobile network signalling data, captured from the continuous interaction of mobile terminals with the cellular network, have better spatial/temporal resolution than traditional Call Detail records (CDR). However, their format and semantic are intimately connected with network-specific technical aspects. For this reason, such data are considerably more complex and have a higher degree of heterogeneity across different Mobile Network Operators (MNO). It is difficult for experts outside the telecommunication domains, such as e.g. statisticians, to interpret and manipulate such data directly. In the proposed contribution we present a general Reference Methodological Framework (RMF) intended to facilitate the use of signalling data by statisticians. The RMF is inspired by the principles of functional layering and by the "hourglass model", which lie at the foundation of modern computer network architectures. The RMF encompasses a convergence layer that decouples the complexity of signalling data at the bottom from the statistical definitions on the top. This allows experts from the two domains, MNO engineers and statisticians, to work independently and eases the evolution of the two layers. This paper presents the general principles underlying the RMF, the role and responsibilities of the different actors in transforming elemental data into meaningful and relevant statistical concepts, provides a concrete actionable proposal and presents early results from its application in a pilot project conducted in collaboration between Eurostat and one European MNO. We highlight lessons learned and give an outlook for the future development and implementation of the RMF and its application to tourism statistics and other areas of statistics.

1 Introduction and motivations

Mobile Network Operator(s) (MNO) collect data from their network infrastructure as part of the normal operation and billing processes. Such data embed information about the activity, location and movements of the mobile devices served by the mobile network. Since almost two decades, researchers in various domains have demonstrated the possibility to exploit MNO data to study human mobility patterns at different spatio-temporal scales. More recently, also MNO have recognised the potential to leverage such data to generate new revenues through mobile analytic services and products, and some of them have established new analytics projects and teams to develop this new line of business. At the same time, also public authorities have become aware of the potential of MNO data for evidence-based policy making.

Historically, the vast majority of research study have considered exclusively Call Detail Records (CDR) [1, 2, 3]. The reason is twofolds. First, CDR are relatively easy to extract from the MNO infrastructure, as they are anyway collected for billing purposes. Second, their structure and semantic can be easily understood by researchers from non-telecom domains, with no need to acquire technical expertise about the mobile network configuration and operation process. In other words, CDR are "low cost" both from the perspective of the data *producer*

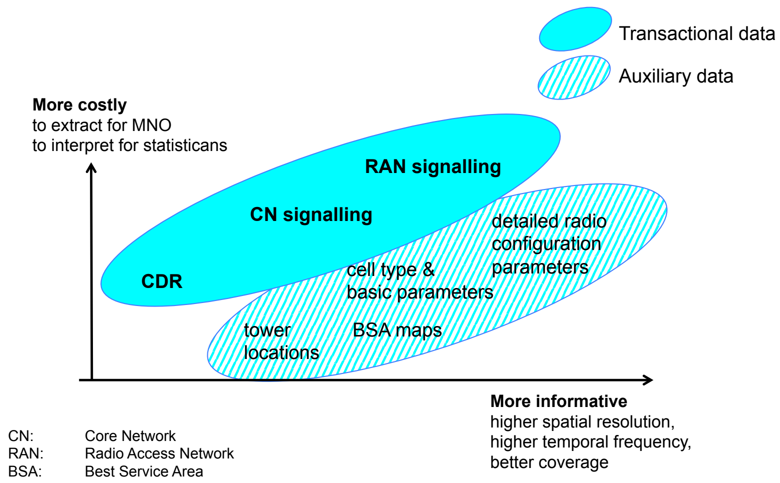


Figure 1: Variety of MNO data types in the D-layer.

(i.e., the MNO staff in charge of the extraction) as well as of the data *consumer* (i.e., the experts from other domains in charge of the analysis).

While CDR is the most basic data source within the MNO domain, it is not the only option. MNO have the possibility to acquire *signalling data* from their networks. Signalling data are more informative than CDR: they have higher frequency in time and potentially also better resolution in space[1]. Furthermore, since signalling events are generated also outside active calls, they are somewhat less exposed than CDR to certain selectivity and bias issues caused by correlations between calling patterns and mobility patterns[2]. On the other hand, extracting signalling data is much more complex than extracting CDR, and typically requires putting in place advanced monitoring systems. Also, their structure and semantic is deeply connected to technical aspects related to the configuration and operation of the mobile network infrastructure. For this reason, a certain degree of technical knowledge is required to interpret and make sense of these data. This clearly represents an additional barrier (and cost) for non-telecom experts willing to work with such data. In other words, signalling data have “higher cost” than CDR both for the data producer as well as for the data user, but at the same time they have potentially *higher value* in terms of more/better information content.

In a scenario where the potential value of signalling data is not evident to the MNO while the extraction costs are significant, it should come to no surprise that research work based on signalling data has remained in the pioneering stage for almost 15 years [4, 5, 6, 7]. However, more recently the increasing awareness of the intrinsic value of mobile analytics data is fostering some MNO to collect *more and better* data from their networks. In parallel, the cost of commercial monitoring equipment is decreasing, also thanks to the evolution of the network architecture (e.g., introduction of IP signalling). The combined effect of these two trends is increasing the interest for mobile analytics based on signalling data as sketched in Figure 1.

From a methodological point of view, the trend towards adoption of signalling data amplifies the need to develop a clear reference framework. Most existing research work in MNO data analysis takes the form of ad-hoc case studies. Each individual study tends to present results obtained with a methodology that was developed ad-hoc for a specific objective, often tailored

[1]This is particularly true for signalling data obtained from the Radio Access Network (RAN).

[2]Note however that signalling data, likewise CDR, remains exposed to other types of bias issues, like e.g. selectivity in mobile customer basis and/or spatial correlations between the radio network layout and mobility patterns.

to the peculiarities of the particular dataset at hand. In principle, the methodology developed ad-hoc for one specific MNO dataset can be generalised to work with other *similar* datasets form other MNO. Moving from CDR to signalling data, we will have to cope with a higher degree of heterogeneity in the characteristics of input dataset from different MNO. This is due to the fact that, as said earlier, the generation of such data is more intimately connected to network-specific configurations and technical aspects. For this reason, ad-hoc methodologies developed for signalling data will be more difficult to port to other dataset, unless *portability* is taken upfront as a design criterion. Besides portability, another limitation of ad-hoc methods relates to their *evolvability*. In fact, the MNO infrastructure keeps evolving continuously, and so do the data format and semantic, especially for what concerns signalling data. An ad-hoc method tailored to a specific dataset might not be easily adapted to work with a future dataset from the same MNO, if a change in the network infrastructure and/or in the monitoring system used to extract the signalling data has occurred meanwhile.

To address these challenges – knowledge gap between telecom engineers and statisticians, portability, evolvability, etc. – EUROSTAT is working towards the definition of a general Reference Methodological Framework (RMF) for the processing of MNO data for Official Statistics. The main objectives of the RMF are:

- Establish a *common terminology* and an effective *conceptual framework* that facilitates cooperation and interworking between experts from the two domains, namely telecom engineers from MNO and statisticians from Statistical Office(s) (SO).
- Ensure a good degree of *generality*, *portability* and *evolvability* of the specific instances of processing and analysis methods developed within the more general framework.
- Provide a concrete basis to clarify the legal aspects connected to the use of MNO data for statistical production, also considering the entry into force of the EU General Data Protection Regulation (GDPR) [8].

With respect to the latter point, we highlight that the development of a unified reference framework, with clear definitions of data types and interfaces at each stage across the workflow, facilitates the process of assessing the related legal aspects as it allows to pose more concrete questions about the compliancy of individual data processing steps. This reduces the room for erroneous and/or over-cautionary interpretations of general legal principles, and ultimately facilitates the dialogue with data protection authorities.

2 General design principles: layering and hourglass model

In the conceptual design of the RMF we borrow some fundamental principles from the field of Computer Networks. Such principles have been the key to the success of the Internet as we know it today [9, 10, 11]. A high-level view of the proposed RMF is represented in Fig. 2. We identify three distinct layers:

- Data layer (**D-layer**) that logically embeds the raw MNO data, at the bottom;
- Statistics layer (**S-layer**) that contains the definitions of statistical indicators, at the top;
- Convergence layer (**C-layer**) that mediates between the D-layer and S-layer.

The computing process runs from the bottom to the top. The data producers (i.e., MNOs) are logically placed in the D-layer, while the entities that "consume" such data for statistics production, i.e., SO, are logically placed in the S-layer (ref. Fig. 2(b)). The D-layer and S-layer map

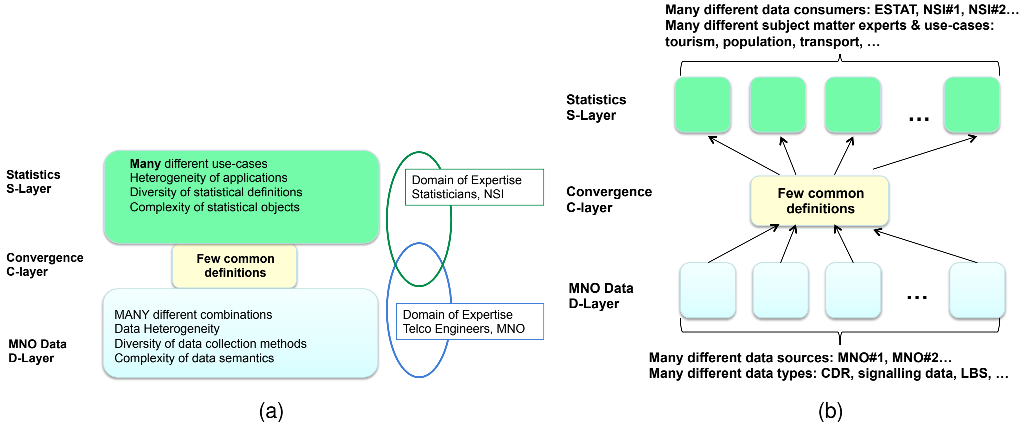


Figure 2: Representation of the layered structure of RMF: the C-layer is the waist of the "hourglass".

naturally to different domains of expertise: telco engineers and MNO experts at the D-layer, statisticians and SO experts at the S-layer. In this context, the C-layer provides the "common language" for the different actors across the two domains.

The D-layer contains a variety of different types of data that can generally grouped into two distinct classes: *transactional data* and *auxiliary data*. Transactional data are generated from the interaction of individual mobile devices with the mobile network infrastructure, therefore contain information about individual mobile users (micro-data). Examples of transactional data are: Call Detail Records (CDR), signalling data from the Core Network (CN), signalling data from the Radio Access Network (RAN), and possibly data from Location Based System (LBS). Auxiliary data relate to the layout and configuration of the mobile network infrastructure: tower locations, radio cell type, radio configuration parameters, Best Service Area (BSA) maps, etc. The combination of transactional data with auxiliary data allow to infer the spatio-temporal mobility patterns of individual mobile devices. The mix of available transactional and auxiliary data varies across different MNO. Even for the same nominal data type (e.g., CN signalling data) the fields actually contained in the records as well as their spatio-temporal resolution depend on the particular (physical and logical) configuration of the network. Generally speaking, there is a trade-off between the "informativeness" and "cost" of each data type, as depicted in Fig. 1.

The S-layer contains a variety of different statistical definitions and indicators that are specific to different use-cases across different subjects (e.g., demography, transport statistics, tourism statistics, etc.). In summary, both D-layer and S-layer are characterised by:

- **Heterogeneity** of the types of data (bottom) and statistics (top).
- **Complexity** of the data structure (bottom) and statistics definitions (top).
- **Multiplicity** of the involved actors: several contributing MNO (bottom) and several statistics domains on the side of SO (top).

In this context, the role of the C-layer is to *decouple* the other two layers, hiding the multiplicity / heterogeneity / complexity of each domain to the other. The C-layer provides a parsimonious common set of structures and definitions that can be understood and processed *independently* by the experts of both domains. The role of MNO experts (telecom engineers) is to produce

C-layer structures from D-layer data (i.e., transactional and auxiliary data): this task is termed “D2C mapping” in Fig. 3. Conversely, the role of SO experts (statisticians and subject matter experts) is to produce unambiguous definitions of S-layer statistics taking C-layer structures as input: this task is termed “C2S processing” in Fig. 3. In this conceptual framework, once the C-layer structures have been defined and accepted by experts from both domains (MNO and SO), the D2C mapping and C2S processing can be developed independently from each other, with no need of continuous and close interaction between experts across the two domains. Furthermore, future changes in the configuration of the MNO data at the D-layer will not invalidate the S-layer definitions already developed, and vice-versa, the modification of existing definitions at the S-layer (or the additions of new statistics) will not require changes to the underlying D2C processing instances already in place. In other words, the C-layer breaks inter-dependencies and allow experts from the two domains to design their respective parts of the processing workflow in a more independent way.

The data structures at the C-layer must be defined according to the following criteria:

- **Parsimony**: few structures, not many;
- **Clarity**: their semantic should be understandable and accepted by experts of both domains (telecom engineers and statisticians);
- **Feasibility**: they should represent (encode) what can be realistically obtained from the underlying MNO data;
- **Sufficiency**: they should retain from MNO data at the bottom the whole amount of potentially useful information for the upper statistics. In other words, no useful information should be lost through the mapping from D-data to C-structures.
- **Generality**: they should not be tailored to the specific characteristics of a particular MNO dataset or infrastructure, nor the specific requirements of a particular use-case. Instead, it should provide support for evolving MNO configuration (at D-layer) and addition of new statistics use-cases (at S-layer) without requiring changes to the C-layer.

Another important design aspect relates to the distinction between *design* and *execution* of the processing method (algorithm). The layering structure described insofar, with the segmentation between D2C and C2S functions, is relevant to the design phase (Who designs each processing module). Also the processing execution (Where the processing module is run) can be segmented between the two organizations, with part of the computation taking place within the MNO domain and another part in the SO, but the segmentation in execution is not bound to mirror the segmentation in design. This concept is exemplified in Fig. 4. Note that C-layer structures still encode personal micro-data, while the reduction from micro-data to macro-data– through the composition of elementary primitives like selection, aggregation etc. – takes place entirely in the S-layer. Therefore, a possible strategy would be to host within the MNO domain the execution of the initial part of C2S functions (that were logically designed by statisticians), and then pass the intermediate aggregate data (non-personal) to the SO, where the final part of the C2S processing workflow is executed. In this way the C-layer micro-data remain fully within the MNO administrative domain, with no transfer of personal micro-data from MNO to SO.

3 Review of MNO data

Hereafter we review MNO data sources beyond traditional CDR. We start presenting three types of transactional data and then delve into auxiliary data.

Signalling data from the Core Network (CN-SIG). The Core Network Signalling (CN-SIG) data are produced by passively monitoring the signalling exchanges in the Core Network (CN) section. They are typically extracted by means of proprietary commercial systems with probes (sensors) placed at appropriate interfaces in the CN, and with the ability to interpret and track the message exchange across the 3GPP protocol stack [4, 12]. The information embedded in CN-SIG data is in general a superset of the CDR data. In other words, if complete signalling data are available, CDR data are redundant and therefore can be ignored.

Signalling data from the Radio Access Network (RAN-SIG). The Radio Access Network Signalling (RAN-SIG) data are produced by passively monitoring the signalling exchanges in the Radio Access Network (RAN) section. Likewise CN-SIG, they require the deployment of sophisticated monitoring systems. Generally speaking, RAN-SIG data are more costly to extract than CN-SIG (ref. Fig. 1), but they are also more informative and can potentially increase the resolution of the observed mobility patterns both in time and space. First, certain network events trigger a RAN-SIG message, but not a CN-SIG message: that means, the temporal resolution of the information embedded in RAN-SIG is potentially higher than CN-SIG. Second, from RAN-SIG one might in principle extract additional measurements related to the radio channel between the mobile device and the serving station that, if appropriately processed, can help to narrow down the size of the inferred C-location. For examples, based on so-called Timing Advance (TA) information, the distance (range) between the serving antenna and the mobile device can be approximately estimated (ref. Fig. 5(c)).

Data from proprietary Location Based Services (LBS). Some MNOs adopt proprietary systems to provide Location Based Services (LBS) to their customers. Such systems aim at identifying the position of the mobile device, based on the radio measurements that the mobile terminal gathers from neighbouring stations *during an active connection*. Such measurements are reported to the network, and from there they can be passed to LBS systems. The latter will use multi-lateration techniques to infer the accurate position of the mobile device based on such measurements. The spatial resolution of LBS varies across different commercial systems and Radio Access Technology (2G, 3G or 4G). Furthermore, LBS estimates might be available only for a subset of the mobile users (depending on contractual agreements with the MNO), and only during certain activity periods.

Auxiliary data. In order to map the radio cell information (typically, the Cell-ID) contained in CDR, CN-SIG and RAN-SIG to a geographical area, such data need to be fused with auxiliary data about the configuration of the individual radio cell. The radio cell attributes that are potentially relevant to this task are: Geographical coordinates of the antenna tower; Antenna height, beam width (circular vs. sector), azimuth orientation and tilting; Radio Access Technology GSM-900, GSM-1800, UMTS, LTE) and frequency band; Cell type: macro-cell, micro-cell, pico-cell, femto-cell. Historically, the vast majority of mobility studies based on MNO data have used exclusively the first item, i.e., tower locations, in combination with Voronoi tessellation. Only few previous work have considered the exploitation of additional auxiliary data [5, 6, 7]. principle, different options for mapping signalling data to C-location can be chosen, taking into account different combinations of transactional and auxiliary data and with different levels of sophistication. Some examples are illustrated in Fig. 5. Higher spatial resolution can be obtained at the cost of using more detailed data and with more sophisticated methods. However one must consider that more detailed auxiliary data are not only more costly to extract, but also more costly to maintain: in fact, the radio network is subject to continuous modifications (reconfiguration, optimisation, growth) and the auxiliary data that are passed to the D2C function need to be kept up-to-date. This clearly represents an additional operational burden.

For a given subset of auxiliary data, one might consider different levels of sophistication in the determination of the smallest geographical area that is "likely" to include the (unknown) device position, i.e., in the identification of C-locations. Recall that "radio cell coverage" is not a static concept: the actual cell coverage varies in time depending on a multitude of internal factors (e.g., network load) and external factors (e.g., weather conditions). In the simplest option, MNO experts might define a *static* mapping that ignores such variability and dependencies, targeting the coarse determination of a single "reference coverage area" that, depending on the design choice, might represent a sort of global average or, conversely, a conservative worst-case estimate. At the opposite extreme, one might develop a sophisticated model that takes in input additional data (e.g., weather parameters, network load) in order to gain spatial resolution. Within this range of possibilities, the optimal choice as to *(i)* which auxiliary data to consider and *(ii)* which mapping model to use (static or dynamic) to use is a matter of balancing the costs (of collecting and maintaining additional data, and of developing the spatial mapping model) vs. benefits (of improved spatial resolution). For initial pilot projects, it is reasonable to expect that the optimal choice will be very close to the simplest one, i.e., static mapping based on nominal parameters. This is fully acceptable at the initial stage. However, as in other realms of technological development, the optimal "costs-vs-benefits" trade-off is a moving target: as MNOs will develop their mobility analytics services, the business value of improved spatial resolution will increase, and the optimal trade-off point will shift towards more sophisticated dynamic solution. This trend is already at play in the shift from CDR towards signalling data sketched in Fig. 1.

4 C-layer structures for MNO data

In this section we detail our initial proposal for the C-layer structures: C-location, C-path and C-attributes. The first two relate to spatio-temporal concepts and refer to the instantaneous location and trajectory of a generic mobile device *as can be inferred from the available MNO data*. The latter refers to a set of mobile device attributes derived from the signalling data.

Before progressing further, it is useful to remind us of the distinction between the (unknown) *actual mobility* and *observed mobility* — as observed by the MNO infrastructure. Generally speaking, we wish to estimate *statistical indicators for actual mobility* based on the *available data related to observed mobility*. The gap between observed and actual mobility relates primarily to the unavoidable finite resolution in space and discontinouity in time. In space, we cannot observe precisely the actual point-position of the generic mobile device, but only identify some extended area that likely contains the (unknown) actual position with a certain high probability. In time, we cannot observe such area continuously at any instant, but only at some discrete set of observation instants (corresponding to the timestamps associated to the signalling events recorded in the signalling data). Furthermore, we should watch against risks of selectivity and correlations affecting the observed mobility process that might eventually translate into bias in the final statistics. In order to grasp the fundamental difference between *actual mobility* and *observed mobility* we may refer, by analogy, to the relation between a physical object and its image captured by low-quality camera. Such analogy is useful to distinguish the various sources of uncertainty that affect the sensing process: lack of resolution (out of focus), missing data (dark pixels), noise, bias (lens distortions).

C-location. For a generic mobile device and observation instant the C-location represents the geographical area that is likely to contain the device, as can be inferred from the available data at the D-layer. The determination of the C-location depends on the particular combination of transactional and auxiliary data that are made available by the MNO. In principle, the C-location might map to the Voronoi cell associated to the tower location, but this approach, while very common in the literature, is rather simplistic and has important limitations. Some authors have introduced variants to the Voronoi procedure to mitigate some of the know limitations (see e.g. [13, 14]). More in general, the C-location should be referred to the radio coverage area of the radio cell. The latter can be estimated from the radio cell type and antenna configuration parameters based on more or less sophisticated propagation models as done e.g. in [6, 7, 15, 16]. Examples of C-locations are illustrated in Fig. 5.

C-path. For a generic device, the C-path represents the sequence of C-locations at the available observation instants. Furthermore, if signalling data are available at the D-layer, the C-path may optionally include also additional information bounding the (unknown) geographical position taken by the mobile devices between two consecutive observation times[3].

C-attributes. These attributes qualify (but not identify) the individual mobile device, i.e., they indicate the association to certain classes that might be relevant for certain use-cases at the S-layer. C-attributes must be extracted exclusively from the the transactional data in the D-layer (at this stage no correlation with other customer database is foreseen). For instance, C-attributes can encode the classification into home subscribers vs. roamers, and in the latter case also the country of origin (this information is readily available from IMSI prefix). Also, it can encode different classes of customers (business contracts, prepaid, etc.) and type of terminal (e.g., smartphone vs tablet).

5 Outlook on ongoing work

We have presented the general principles and initial ideas about a novel Reference Methodological Framework for the processing of MNO data for Official Statistics. The ideas and concepts presented in this work represent the intermediate (preliminary) results from an ongoing strand of work in EUROSTAT. The work benefits from a continuous dialogue with technical experts from Proximus, the leading MNO in Belgium, and with other National Statistical Institutes within the European Statistical System. In the future progress of this work we will propose more detailed specifications and implementation guidelines about the RMF, particularly for what concerns C-layer structures and D2C mapping functions. We also plan to develop concrete instances of processing methodologies, in agreement with the general RMF outlined here, for a small number of concrete use-cases in selected application domains – most prominently density estimation and tourism – based on real-world data. Furthermore, besides the technical and methodological aspects, we are also committed to continue the dialogue with other stakeholders to clarify other aspects related to legal compliancy, business models and processes [17].

References

[1] M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196), 2008.

[3]This information can be derived from signalling data and relates to the notions of Location Area (in 2G), Routing Area (in 3G) and Tracking Area (in 4G).

[2] K. S. Kung, K. Greco, S. Sobolevsky, and C. Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS ONE*, 9(6), 2014.

[3] T. Louail, M. Lenormand, O. Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. Ramasco, and M. Barthelemy. From mobile phone data to the spatial structure of cities. *Nature Scientific Reports*, 4(5276), 2014.

[4] F. Ricciato et al. Traffic monitoring and analysis in 3g networks: lessons learned from the metawin project. *Elektrotechnik & Informationstechnik (in English)*, 123/7/8, 2006. http://ricciato.unisalento.it/E&I2006_pub-et_11366.pdf.

[5] P. Fiadino, D. Valerio, F. Ricciato, and K. Hummel. Steps towards the extraction of vehicular mobility patterns from 3g signaling data. In *4th Int'l Workshop on Traffic Monitoring and Analysis (TMA 2012), LNCS 7189*, March 2012.

[6] A. Janecek, D. Valerio, K. Hummel, F. Ricciato, and H. Hlavacs. The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE Transactions on Intelligent Transportation Systems*, 16(5), October 2015.

[7] F. Ricciato, P. Widhalm, M. Craglia, and F. Pantisano. Beyond the “single-operator, cdr-only” paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing*, May 2016.

[8] EU General Data Protection Regulation (GDPR), March 2018. https://www.eugdpr.org.

[9] S. Akhshabi and C Dovrolis. The evolution of layered protocol stacks leads to an hourglass-shaped architecture. In *ACM SIGCOMM'11*, August 2011.

[10] M. Chiang, S. Low, A. Calderbank, and J. Doyle. Layering as optimization decomposition. *Proceedings of the IEEE*, 95, 2007.

[11] J. Doyle. The architecture of robust, evolvable networks. *The Lee Center for Advanced Networking*, Spring 2010. https://tinyurl.com/y7z7haze.

[12] F. Ricciato. Traffic monitoring and analysis for the optimization of a 3g network — special issue on 3g/4g/wlan/wman planning. *IEEE Wireless Communications*, 13(6), Dec. 2006.

[13] Center for Spatial Information Science — Univ. of Tokyo. A study on urban mobility and dynamic population estimation by using aggregate mobile phone sources. http://www.csis.u-tokyo.ac.jp/dp/115.pdf.

[14] F. De Meersman et al. Assessing the quality of mobile phone data as a source of statistics. In *European Conference on Quality in Official Statistics*, June 2016. https://tinyurl.com/y7pbracn.

[15] F. Ricciato, P. Widhalm, M. Craglia, and F. Pantisano. Estimating population density distribution from network-based mobile phone data. JRC Technical Report, 2015. https://tinyurl.com/ydz4mgaw.

[16] M. Tennekes. R package for mobile location algorithms and tools, April 2017. https://github.com/MobilePhoneESSnetBigData/mobloc.

[17] F. Ricciato, F. De Meersman, A. Wirthmann, G. Seynaeve, and M. Skaliotis. Processing of mobile network operator data for official statistics: the case for public-private partnerships. In *104th DGINS conference*, Oct. 2018. https://tinyurl.com/yahwq3dm.

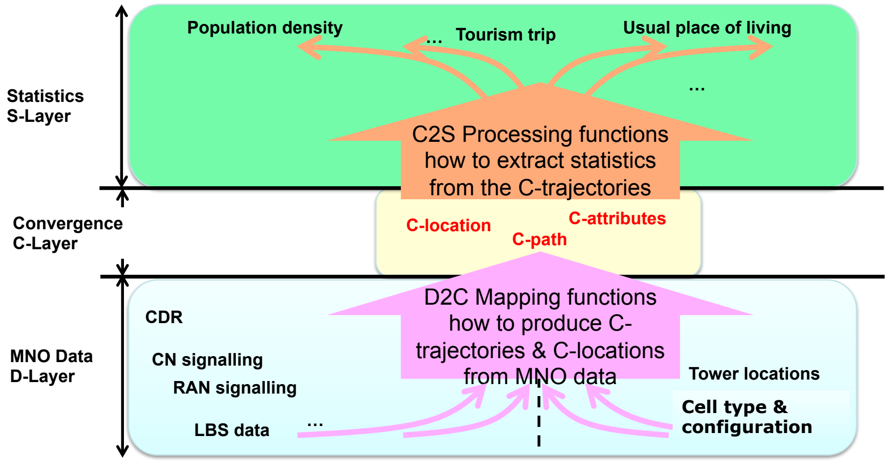


Figure 3: Decoupling of D2C and C2S functions.

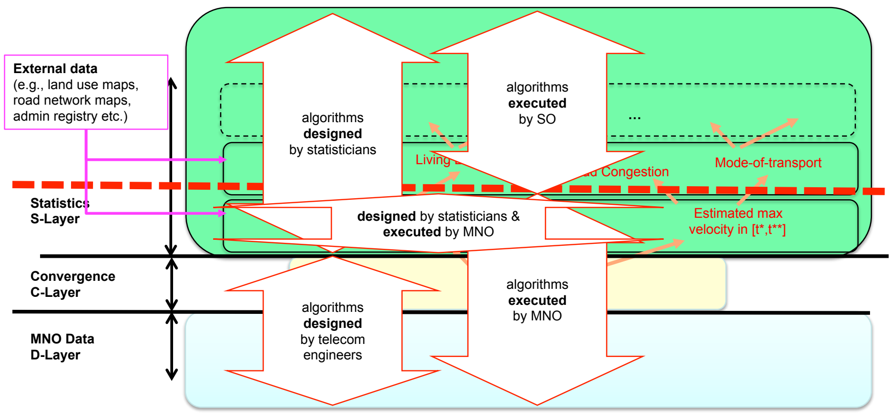
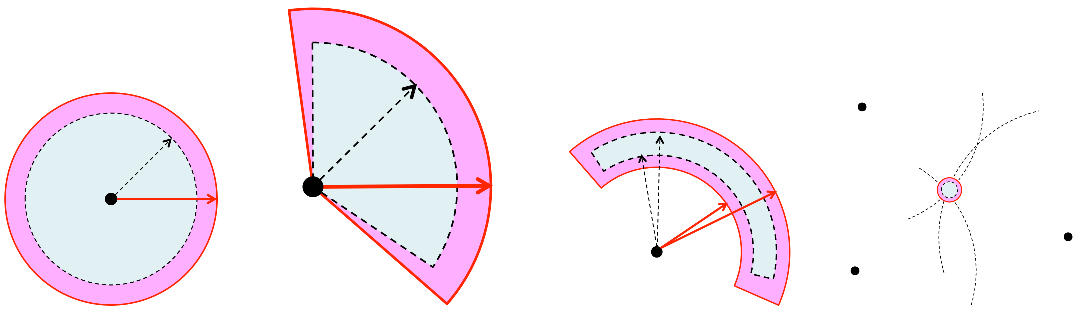


Figure 4: Segmentation in the processing design vs. execution. The bottom part of the S-layer still includes individual micro-data and is executed within the MNO domain. Intermediate aggregate data (non-personal) that are produced within the S-layer are then exported to the SO.



(a) Circular cell

(b) $120°$ Sector

(c) $120°$ Sector with range information (TA)

(d) Location Based Systems estimate

Figure 5: Examples of C-locations (magenta) inferred for D-layer data.