# Towards a Reference Methodological Framework for processing Mobile Network Operator data for Official Statistics

Fabio Ricciato*, Albrecht Wirthmann*,
Martijn Tennekes†, Benjamin Sakarovitch‡, Roberta Radini§ and David Salgado¶

*EUROSTAT European Commission.
†CBS, Netherlands. ‡INSEE, France. §ISTAT, Italy. ¶INE, Spain.

An increasing number of Mobile Network Operators (MNO) are now able to extract signalling data originating by the continuous interaction of mobile stations with the cellular network. Compared to traditional Call Detail Records (CDR), signalling data have better spatial/temporal resolution and represent a more informative data source to analyse human presence and mobility patterns. On the other hand, such data are considerably more costly to extract and more complex to interpret than CDR. Furthermore, they yield a higher degree of heterogeneity across different MNOs, due to proprietary monitoring technologies and tighter dependencies with the underlying network configuration. Their format and semantics change, following the evolution of MNO infrastructure. Consequently, it is difficult for statisticians and researchers without a solid telco engineering background to interpret such data.

In order to overcome these challenges, and thus promote the adoption of MNO signalling data for the production of official statistics, we propose a general Reference Methodological Framework (RMF) intended to facilitate the use of signalling data by statisticians and, in general, by non-telco experts. The RMF is inspired by the principles of modularity, functional layering and the "hourglass model" that lie at the foundation of modern computer network architectures.

Broadly speaking, the RMF encompasses three functional layers: an intermediate convergence layer (C-layer) decouples the complexity of signalling data at the bottom (D-layer) from the statistical definitions on the top (S-layer). Decoupling the D-layer and S-layers allows experts from the two domains, namely MNO engineers and statisticians, to work independently, easing the evolution of both layers. In the proposed vision, MNO engineers implement low-level data processing functions, collectively called "D-to-C" mapping, that transform raw network data into a sequence of geo-located events in a common format, along with additional auxiliary information, that represent C-layer data. The proposed model can accommodate different geo-location methods for individual events, including simplistic Voronoi tessellation (a popular but sub-optimal choice) and more advanced approaches based on cell radio coverage predictions.

From the C-layer data, the statisticians develop estimation methods that are logically placed in the upper S-layer. In so doing, they might fuse geo-located C-layer event data with external geographical data, e.g. transportation maps and/or land use maps.

Adoption of a common RMF enables the reuse of algorithms and processing modules developed by different statistical offices and/or research groups across data from different MNO. In this sense, it will facilitate benchmarking, independent validation and collaborative development of processing algorithms developed by academic and industrial research groups. When coupled with open-source algorithms, it helps to achieve transparency, auditability and public scrutiny onto processing algorithms.

The RMF is being developed within the European Statistical System by a joint effort between EUROSTAT and the National Statistical Institutes of some European countries. In this presentation we will present the general principles underlying the RMF, report on the status of ongoing development and outline future work.