ESTAT 2019.0232 Technical Note

Specification of test use-cases

Fabio Ricciato

Eurostat – Unit A5 Methodology; Innovation in official statistics

Email: fabio.ricciato@ec.europa.eu

Abstract

This is a technical document prepared by Eurostat. It specifies the test use-case to be implemented by Cybernetica in the context of the Eurostat-Cybernetica project ESTAT 2019.0232.

The data processing logic defined in this document serves exclusively the purposes of testing and demonstrating the feasibility of the solution developed in this specific project. As such, it constitutes a statistical "toy methodology" and adopts statistical "toy definitions" that are valid only within this scope of this experimental project, and should not be considered representative of any (present or future) official methodologies or definitions.

I. Description of input data

The (raw) event data during a period T (by default, $T = 24$ hours) are pre-processed by the MNO-ND through module "A" in Fig. 2 that runs outside the Cybernetica solution. In the pre-processing stage, the data for the generic mobile subscriber $m = 1, \ldots, M$ are processed independently from other subscribers. The total number of mobile subscribers for testing is to be set to $M = 5 \cdot 10^7$ (50 millions) or higher. Let $p_{k,[m]}$ denote the pseudonym associated to mobile subscriber m during interval k.

For the generic interval k, the pre-processing function takes in input the batch of events associated to the same pseudonym $p_{k,[m]}$ and returns in output a data structure $\boldsymbol{H}_{k,[m]}$. For the sake of illustration, in this document we consider $\boldsymbol{H}_{k,[m]}$ to be a sparse matrix, but of course any suitable data structure (vector, list...) may be used in the implementation.

The matrix $\boldsymbol{H}_{k,[m]}$ has size $(I_T + 1) \times J$, with $J \approx 10^6$ and $I_T = 3$, and it is *highly sparse*. The elements of this matrix can be assumed to take integer values. This matrix represents the

"intra-period footprint" during the observation interval k of the generic mobile subscriber m (with pseudonym $p_{k,[m]}$). The generic element (i, j) of this matrix represents a score[1] assigned to grid unit j in the daily sub-period i, as defined hereafter with support of Fig. 1.

The national territory is divided into a regular grid of 1 km $\times$ 1 km. Additional (auxiliary) grid units are considered as representatives of foreign mobile networks (for outbound roamers). All grid units (regular and auxiliary) are indexed in $j = 1, \ldots, J$. For the project, we shall assume $J = 10^6$ as a conservative choice[2]. I will use the term "tile" as synonymous for "grid unit" (note[3]).

The 24h daily cycle is divided into $I_T = 3$ daily sub-periods, not necessarily disjoint, corresponding to night time, working time and evening. For each grid unit j, we consider $I_T + 1$ score elements indexed in $i = 0, \ldots, I_T$: the first element refers to the whole period T, while the remaining I_T elements refer to the different daily sub-periods, as graphically depicted in Fig. 1

II. Aggregation

A. Overview and modular structure of the processing flow

The processing that will take place in the TEE can be divided into a chain of separate blocks labelled "B", "C" and "D" in Fig. 2.

The matrix $\boldsymbol{H}_{k,[m]}$ is computed (by MNO-ND) on the pseudonymised raw data and represents and *input to the Cybernetica solution* (based on Trusted Execution Environment, TEE) together with the pseudonym $p_{k,[m]}$.

B. Accumulation of individual footprint

The module "B" (first module in the Cybernetica solution) performs for each individual mobile subscriber m a simple *additive aggregation* of the *intra-period footprint* matrices collected until time τ. In this way it builds the *longitudinal footprint* $\boldsymbol{S}_{\tau,m} \stackrel{\text{def}}{=} \sum_{\tau=1}^{k} \boldsymbol{H}_{k,[m]}$ for each

[1]The exact interpretation of such "score", as well as the detailed algorithm used to compute it from a sequence of individual events, are not relevant for the purpose of the Eurostat-cybernetica project, since they related to the pre-processing function in module "A" that is outside the cybernetica solution. The detailed definition of such "score" and the algorithmic logic of module "A" are the focus of a separate parallel project.

[2]The area of a large EU country like France has about 600,000 km^2.

[3]we shall avoid the use of term "cell" for this project, since I would like to keep this term reserved to denote "radio cells" in the raw MNO data.

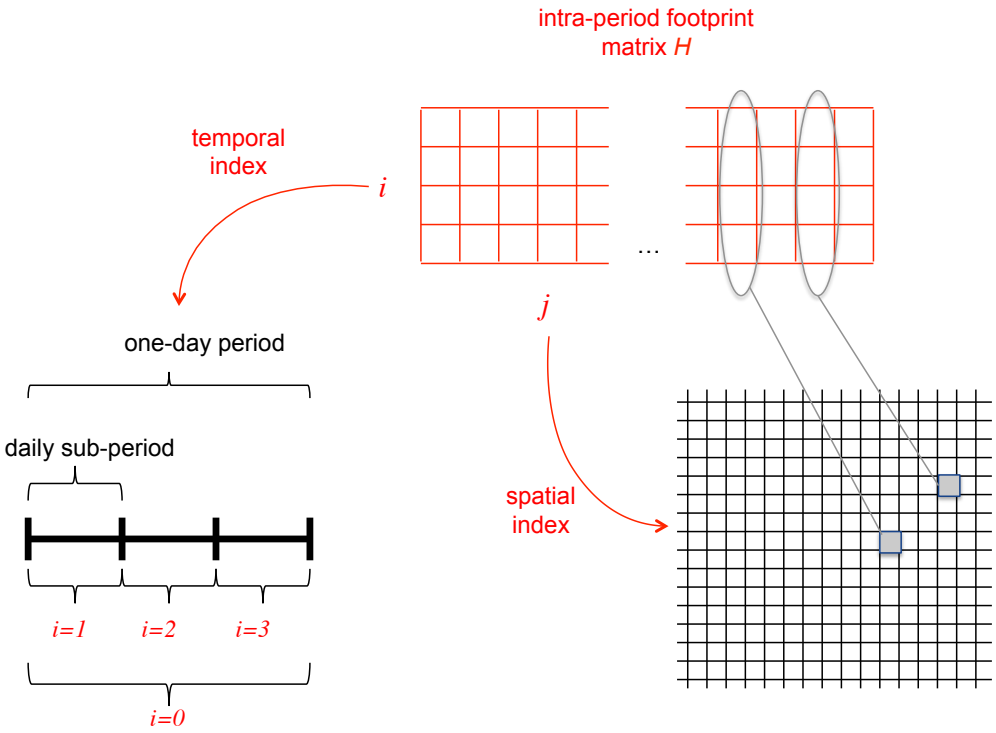


Figure 1. Interpretation of intra-period footprint $\boldsymbol{s}$.

individual mobile subscriber m. This corresponds to a simple temporal integration that can be also implemented sequentially as:

$$\boldsymbol{S}_{k,m} = \boldsymbol{S}_{k-1,m} + \boldsymbol{H}_{k,[m]} \quad \text{for} \quad k = 1, \ldots, \tau. \tag{1}$$

The main challenge here is to handle the large size of the input and output matrices $\boldsymbol{H}$ and $\boldsymbol{S}$, and the technical solution adopted by Cybernetica will probably leverage the high level of sparsity of both matrices.

C. Consolidation and summarisation of individual accumulated footprint

The next module "C" is run only at time τ. For each subscriber m, it takes in input the longitudinal footprint $\boldsymbol{S}_{\tau,m}$ up to time τ, applies a summarisation function $g_1()$ or $g_2()$, and returns in output a summary structure $\boldsymbol{Y}_m$ or $\boldsymbol{L}_m$, formally:

$$g_1 : \ \boldsymbol{S}_{\tau,m} \rightarrow \boldsymbol{Y}_m \tag{2}$$

$$g_2 : \ \boldsymbol{S}_{\tau,m} \rightarrow \boldsymbol{L}_m \tag{3}$$

(from now onward we omit the index τ). The summarization function might take in input external parameters that, however, are fixed and known at code compilation time.

For this demonstrative project, we assume the summary structure $\boldsymbol{Y}_m$ consists of a binary vector of same size as the longitudinal footprint $\boldsymbol{S}_{\tau,m}$, i.e., $(I_T+1)\times J$. In the first row (corresponding to sub-period index $i=0$) the jth element $y_m(0,j)\in\{0,1\}$ indicates whether tile j is part of the "usual environment" of mobile subscriber m. The summary function can be written as a simple threshold-based quantisation on the first row of the accumulated footprint:

$$y_m(0,j)=\begin{cases}1 & \text{if } S_{m,k}(0,j)\geq\psi\,,\\ 0 & \text{otherwise.}\end{cases} \tag{4}$$

(recall that $i=0$ corresponds to whole-period score). A reasonable value for the threshold is e.g. $\psi=0.3$.

For each of the remaining rows $i=1,\ldots,I_T$ (corresponding to the different daily sub-periods), and only for the tiles included in the usual environment (for which $y_m(0,j)=1$), the jth element $y_m(i,j)\in\{0,1\}$ indicates whether location j was visited "prevalently" in sub-period i. All sub-periods with a relative score higher than a threshold value ϕ are deemed to be "prevalent" (in other words, there might multiple prevalent subperiods, or none). Therefore, the summary function can be written as a simple threshold-based quantisation:

$$y_m(i,j)=\begin{cases}1 & \text{if } S_{m,k}(0,j)\geq\psi \text{ and } \frac{S_{m,k}(i,j)}{S_{m,k}(0,j)}\geq\phi\,,\\ 0 & \text{otherwise.}\end{cases}\qquad i\in 1,2,3 \tag{5}$$

A reasonable setting is $\phi=0.5$.

Note that for some "highly nomadic" mobile subscribers none of the observed tiles fulfill the condition (4) (i.e., $S_{m,k}(0,j)\geq\psi$). For these subscribers the structure $\boldsymbol{Y}_m$ will be empty and they will not be considered in the remaining part of the analysis. However, the analysis code should report the total number of "highly nomadic" mobile subscribers in a separate counter.

The summary structure $\boldsymbol{L}_m$ represents a short list of anchor tiles for the mobile subscriber m, aiming at capturing the main places of living and work/study. Conceptually, it may be considered as an ordered list of tiles. The length of this list is limited to a maximum of L_{max} tiles. (Note: in practical applications, I expect to set the value of L_{max} to a small number. Although for the use-cases tested in this project we will be using only the 1st ranked tile, I suggest to implement the code by setting $L_{max}=4$).

To build such list for a generic mobile subscriber m we proceed as follows:

1 - Consider only the tiles that fulfill the condition (4), for which $S_{m,k}(0,j) \geq \psi$ or equivalently $y_m(0,j) = 1$.

2 Rank them according to the value of $S_{m,k}(0,j)$ in sub-period 0.

3 In case that two or more tiles have same tie score value in sub-period 0 (e.g. j_1 and j_2 such that $S_{m,k}(0,j_1) = S_{m,k}(0,j_2)$) then they will be ranked according to the highest prevalence score in the other sub-periods $\max_{i\in\{1,2,3\}} S_{m,k}(i,j\ )$.

4 In case that also the latter score leads to a tie, then they will be ranked according to highest score in sub-period 1 (corresponding to night time) $S_{m,k}(1,j)$).

5 In case that also this criterion leads to a tie, then they will be ordered randomly (NB: in the design of a random assignment, we should avoid systematic preference of one tile across different mobile subscribers).

6 The top L_{max} tiles will be selected to enter the ordered structure $\boldsymbol{L}_m$ (note we should keep the order!) (note also that L_{max} is a maximum limit, and in some case the number of selected tiles fulfilling the condition (4) could be lower than the limit.)

Note that the set of summaries $\boldsymbol{Y}_m$ and $\boldsymbol{L}_m$ will NOT be revealed to the output parties. However, such values should be kept stored in protected form, accessible only by the enclave, since they will serve as (private) input for the weight calculation process described in section IV.

D. Final aggregation

The final module "D" takes in input the set of summaries for all mobile subscribers $\{\boldsymbol{Y}_m\}_{m=1,\ldots,M}$ and possibly some external parameters, and delivers in output some aggregated statistics. Formally:

$$f:\ \big(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_M, \text{parameters}\big) \rightarrow \text{statistics} \tag{6}$$

We shall consider two main use-cases:

- **Use case #1**. The external parameters are all *public* (or, as a special case, absent).
- **Use case #2**. Some external parameters are *private data* held by the NSI. The NSI makes these data available for the computation but does not disclose them to the MNO.

III. USE CASE #1 WITH ALL PUBLIC PARAMETERS

For this project we consider a very simple final statistics (simple addition) plus a more elaborated one.

The first statistics is merely the sum matrix

$$\boldsymbol{D} = \sum_m \boldsymbol{Y}_m \tag{7}$$

of the subscriber footprints in each temporal sub-periods.

Another companion statistics is obtained by aggregating the mobile subscribers based on their observed top location, as encoded in the ordered list $\boldsymbol{L}_m$. If we assume that the 1st ranked tile in the list $\boldsymbol{L}_m$ corresponds to the "main place of living" of the generic mobile subscriber m, we can count for every tile j the number of mobile subscribers p_j having their "main place of living" in j. Stacking the p_j's for all tiles we build the (sparse) vector $\boldsymbol{P}$ that represents the spatial distribution of observed mobile subscribers by their "main place of living". The vector $\boldsymbol{P}$ will represent an output statistics together with matrix $\boldsymbol{D}$.

NB: For Statistical Disclosure Control (SDC) reasons, all element of $\boldsymbol{D}$ and $\boldsymbol{P}$ below a given threshold ξ (public parameter) are either omitted or merged with neighboring elements (so that the combined sum exceeds the threshold) in the reported output. Let $\boldsymbol{D}'$ and $\boldsymbol{P}'$ denote the resulting value of $\boldsymbol{D}$ and $\boldsymbol{P}$, respectively, after running such SDC operations:

$$\text{SDC}: \quad \boldsymbol{D} \to \boldsymbol{D}' \quad \text{and} \quad \boldsymbol{P} \to \boldsymbol{P}'$$

While $\boldsymbol{D}'$ and $\boldsymbol{P}'$ (after SDC) can be revealed to the output parties, the complete value of $\boldsymbol{D}$ and $\boldsymbol{P}$ (before SDC) should remain stored in protected form, accessible only by the enclave, since it will serve as (private) input for the calibration process described in section IV.

The second statistics has a more elaborated definition inspired by the concept of Functional Urban Area[4] and variants thereof[5]. To avoid confusion with other definitions, we resort to the term "Functional Urban Fingerprint" (FUF for short) to refer to the novel definition defined

[4]See https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Functional_urban_area for a formal definition

[5]See e.g. the approximation considered by the Spanish statistical office in a recent exercise https://ine.es/censos2021/movilidad_proyecto.pdf (FUA), and the alternative concept "Mobility Functional Area" considered in the recent study by JRC https://ec.europa.eu/jrc/en/publication/mapping-mobility-functional-areas-mfa-using-mobile-positioning-data-inform-covid-19-policies.

in this document. The FUF[6] concept is meant to emulate (or roughly approximate) the official notion of FUA within the scope of what is realistically achievable by MNO data and relevant for this specific project[7].

The input parameters for FUF computation are a set of "reference areas" (RA) that are pre-defined by the analyst and roughly correspond to administrative urban areas. The RA parameters are static can be provided at compilation time. Two examples of RA are depicted in Fig. 3. A RA consists of multiple adjacent grid units. Different RAs are expected to be mutually disjoint, but their union does not necessarily coincide with the whole country. We may assume that the total number of RAs given in input is in the order of $N_R = 100$.

Let $\mathcal{A}_r$ denote the geographic scope of the generic RA r ($r = 1, \ldots, N_R$), i.e., the set of tiles included in the RA r (alternatively, the RA scope can be encoded in a binary vector $\boldsymbol{A}_r$ of size J whereby the jth element indicates whether or not tile j is included in the RA).

Let the binary variable $\delta_m(r) \in \{0, 1\}$ indicate whether the usual environment of mobile subscriber m, encoded in the vector $\boldsymbol{Y}_m(0)$ for sub-period 0, has at least one tile in common with the rth RA. This condition can be expressed in multiple ways, e.g. as

$$\delta_m(r) \stackrel{\text{def}}{=} \max_{j' \in \mathcal{A}_r} y_m(0, j') \; = \max\left(\boldsymbol{Y}_m(0) \odot \boldsymbol{A}_r\right) = \begin{cases} 1 & \text{if } \boldsymbol{Y}_m(0) \cap \boldsymbol{A}_r \neq 0 \;\; , \\ 0 & \text{if } \boldsymbol{Y}_m(0) \cap \boldsymbol{A}_r = 0 \;\; . \end{cases}$$

whereby '$\odot$' denotes element-wise product. For a generic RA r and a generic tile $j \notin \mathcal{A}_r$ *outside* it scope, we define the "connection strength" $C(j, r)$ as the ratio

$$C(j, r) = \frac{\sum_m \delta_m(r) y_m(0, j)}{\sum_m y_m(0, j)} \tag{8}$$

The denominator represents the number of mobile subscribers that have tile j in their usual environment. The numerator represents the number of mobile subscribers that have *both tile j and RA r* in their usual environment. *NB: For Statistical Disclosure Control (SDC) reasons, the elements (j, r) for which the numerator of equation* (8) *is below a given threshold ξ are omitted.*

Therefore, for each RA r, the collection of tiles with connection strength above a minimum threshold represents the area FUF.

[6]Note the difference between the term "footprint", referring to data structures associated to individual mobile subscribers (hence indexed in m), and "fingerprint", for data structures associated to reference area.

[7]Fabio: to expand motivation and justification

IV. USE CASE #2 WITH PRIVATE PARAMETERS

In this use case, the NSI provides external information (from census grids or from surveys) to *calibrate* the final statistics. The calibration process is based on *individual weights that are computed for each mobile subscriber based on the fusion of MNO and NSI data*. Such weights can be used to refine (calibrate) other statistics.

The external parameter provided by the NSI is a vector $\boldsymbol{\ell} \stackrel{\text{def}}{=} [\ell_1, \ldots, \ell_J]$ of size J. The generic component ℓ_j denotes the absolute number of residents in tile j according to the census grid held by the NSI. This value is compared with the number p_j of mobile subscribers that, based on the MNO data, are presumed to live "mainly" in tile j.

For a generic tile, we define the adjustment parameter a_j computed as follows:

$$a_j \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } \max\left(\ell_j, p_j\right) < 10, \\ 1/5 & \text{if } \max\left(\ell_j, p_j\right) \geq 10 \text{ and } \ell_j/p_j \leq 1/5, \\ 10 & \text{if } \max\left(\ell_j, p_j\right) \geq 10 \text{ and } \ell_j/p_j \geq 10, \\ \ell_j/p_j & \text{if } \max\left(\ell_j, p_j\right) \geq 10 \text{ and } 1/5 < \ell_j/p_j < 10. \end{cases} \tag{9}$$

The total number of observed mobile subscribers, before and after adjustment, shall be reported, namely $z_{obs} \stackrel{\text{def}}{=} \sum_{j=1}^{J} p_j$ and $z_{adj} \stackrel{\text{def}}{=} \sum_{j=1}^{J} a_j p_j$ (these values are not subject to SDC!).

For a generic mobile subscriber m, the individual weight w_m is set equal to the adjustment parameter a_{j_m} computed from (9) in the tile j_m that is ranked at the 1st place in the ordered list $\boldsymbol{L}_m$.

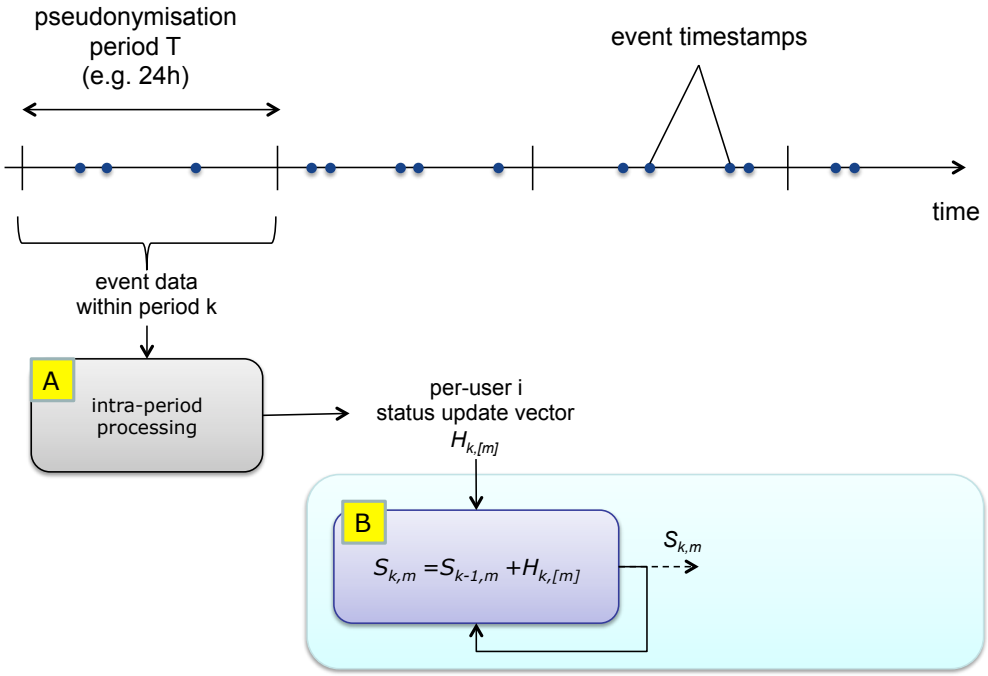
Finally, after computation of the (private) individual weights, the new *calibrated* versions of the previously defined statistics (7) and (8) are computed, formally:

$$\boldsymbol{D}_w = \sum_m w_m \, \boldsymbol{Y}_m \tag{10}$$

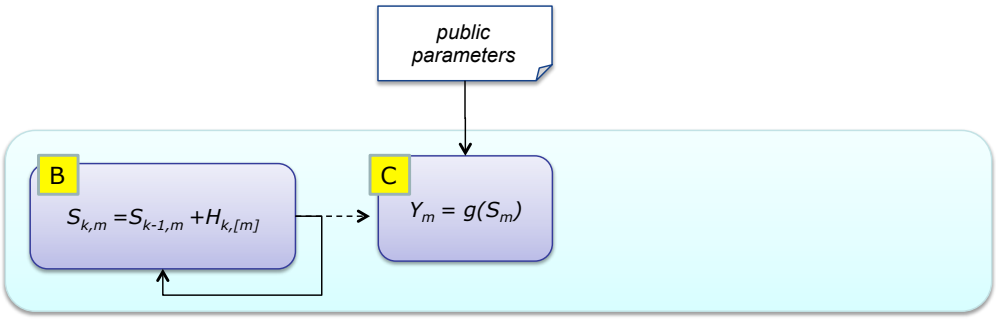
and

$$C_w(j, r) = \frac{\sum_m w_m \, \delta_m(r) \, y_m(0, j).}{\sum_m w_m \, y_m(0, j)}. \tag{11}$$

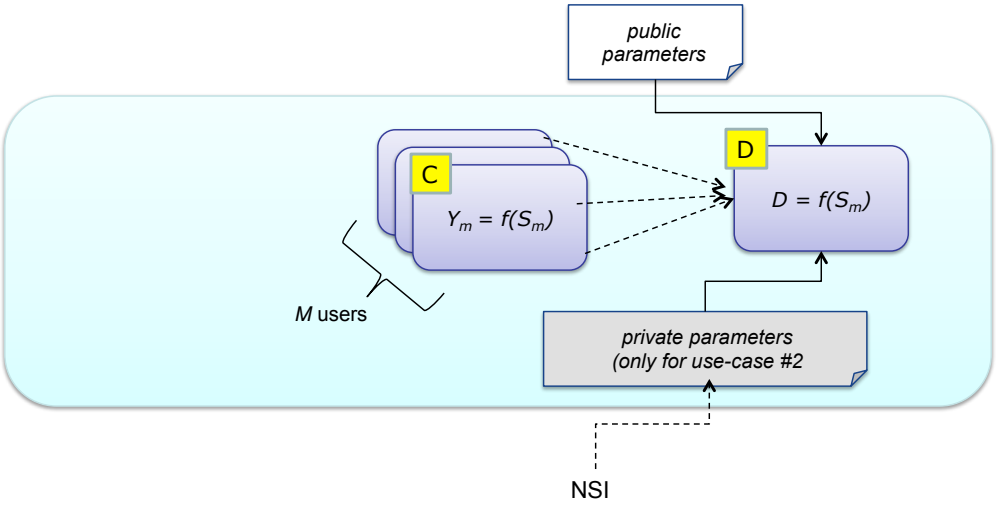
Comparing $\boldsymbol{D}_w$ and $C_w(j, r)$ in (10) and (11), respectively, with $\boldsymbol{D}$ and $C(j, r)$ in (7) and (8), we note the inclusion of individual weights w_m.



(a) Individual integration



(b) Individual summarisation



(c) Collective Aggregation

Figure 2. Schematic flow: the private computation blocks are embedded in the cyan box.

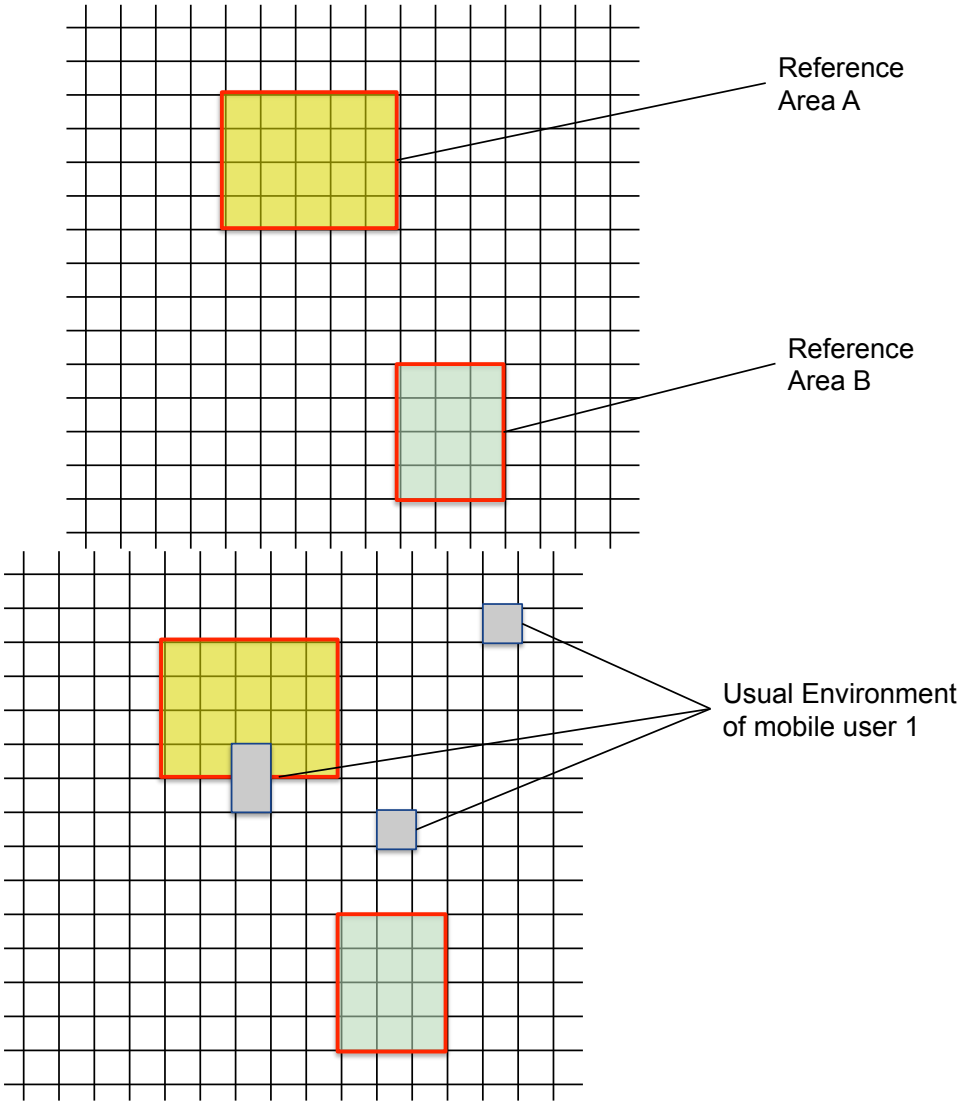


Figure 3. Example of reference areas (RA).