Population Density Data API

Algorithm initial proposal

Inputs:

- **UEs connection records** data is received pseudonymized or hashed.
- **Requested Space** is received in the API customers' request.

Process is as follows:

1. **Cleanup**: Repeated records will be deleted and traffic events associated with M2M lines are filtered due to they do not contribute to the process of estimating future population density.

   Output 1 → Cleaned list of UEs connection records

2. **Attribution**: A cell is associated with each user in each time interval.

   a. For mobile lines with persistent records, the last cell in which each user has been connected in each interval (considering intervals of 15 minutes) will be associated.

   b. It is common to have intervals for which a user does not show activity, for them it will be assumed that the user remains in the same cell in which he was last seen active.

   Output 2 → Valid UEs per cell in each time interval

3. **Counting and Aggregation**: A count of users is performed per cell and interval. On this count, a statistical analysis is performed with the elimination of outliers. Records associated with cells with less than K users in each time interval (k – anonymity) are discarded.

   Output 3 → Total number of UEs connected per cell in each time interval, considering privacy.

4. **Spatial indexing** (no personal data is processed): the space will be divided into units (grids), and associating the grids that overlap with its coverage area to each cell.

   Output 4 → Regular grid covering the required area, over the cells coverage areas.

5. **Distribution and aggregation**: Taking into account the coverage area of the cell, users are going to be distributed, homogeneously among the grids (currently 150m x 150m or larger) associated with the cell, and the process is repeated for all cells. As we are assuming that the distribution of users in the coverage area

of each cell is uniform, we are introducing an error/noise in the distribution of users by time interval, which will be transferred to population density predictions that contribute to reducing the risk of user reidentification. Each grid on the map can typically be served by several cells of different technologies and frequency bands. To obtain the number of users per grid and time interval, an aggregation is performed.

Output 5 → Number of UEs per grid, based on distribution in the cells covering that area

6. **Prediction**: Based on the historical information of users by grid and interval, the prediction of users by grid and interval is made for each time interval of the future. For example, the population density forecast for next Tuesday should be similar to the one observed last Tuesday. As the data becomes available, improvements to this process will be proposed to consider seasonality, holidays/working days and also to be able to make revisions for short-term predictions. For example, on a typical Tuesday we observe a population density of X, but today Tuesday at 10 o'clock we are already 25% above that level, so it is foreseeable that at 11 or 12 o'clock we will also continue above that level.

   Output 6 → Users per grid, considering historical and calendar data

7. **Extrapolation**: Finally, the users of the MNO mobile network are extrapolated taking into account the market shares by geographical area to obtain the total number of people (users or not of the mobile network with any operator).

   Output 7 → Total population per grid, considering extrapolation towards MNO market share

From step 2 onwards, this is aggregated data that does not contain personal information (this is anonymous data).