



# A Cellular Network Database for Fingerprint Positioning Systems

Donatella Gubiani<sup>1</sup>(✉), Paolo Gallo<sup>2</sup>, Andrea Viel<sup>2,3</sup>, Andrea Dalla Torre<sup>3</sup>,  
and Angelo Montanari<sup>2</sup>

<sup>1</sup> University of Nova Gorica, Nova Gorica, Slovenia  
donatella.gubiani@gmail.com

<sup>2</sup> University of Udine, Udine, Italy

{paolo.gallo,angelo.montanari}@uniud.it

<sup>3</sup> u-blox Italia SpA, Trieste, Italy

{andrea.viel, andrea.dallatorre}@ublox.com

**Abstract.** Besides being a fundamental infrastructure for communication, cellular networks are increasingly exploited for positioning via signal fingerprinting. Here, we focus on cellular signal fingerprinting, where an accurate and comprehensive knowledge of the network is fundamental. We propose an original multilevel database for cellular networks, which can be automatically updated with new fingerprint measurements and makes it possible to execute a number of meaningful analyses. In particular, it allows one to monitor the distribution of cellular networks over countries, to determine the density of cells in different areas, and to detect inconsistencies in fingerprint observations.

**Keywords:** Cellular network · Signal fingerprinting · Multilevel database · Data analysis

## 1 Introduction

Nowadays, our society is characterized by a pervasive use of mobile devices. The most common example of a mobile device is the smartphone, which combines voice communication with data services, Wi-Fi connection, and localization services to support advanced activities. As a matter of fact, many commonly employed applications make use of the current position of the user.

In order to compute the current location of a device, the most widely known solution is the Global Positioning System (GPS). Despite its widespread use, GPS has some significant drawbacks. On the one hand, the GPS signal cannot be received in certain conditions; on the other hand, energy consumption of GPS modules can be a problem with battery-powered devices [8, 15].

Cellular signal fingerprinting offers a viable alternative to GPS solutions [1, 2, 7]: an estimation of the current position of a device can be obtained by comparing the signals received at that position with those recorded in a database of

observations taken at known positions. The use of signals coming from cellular networks in positioning systems has a number of advantages. The most significant ones are high coverage and low cost. Moreover, monitored devices need to be equipped with a cellular module and a simple software component only.

The most critical aspect is the need of collecting and maintaining a large set of fingerprints with their position. As a matter of fact, there are some ready-to-use repositories. Some of them are free of charge, others require a subscription fee. The most famous community dataset is OpenCellID [13], which is the result of a crowd source effort. We will make use of such a dataset. Unfortunately, as it happens with the other publicly-available datasets, collected data are poorly structured: information is recorded in a raw table (in csv format) of cellular signal readings paired with their position in a global reference system. In particular, no data structuring reflecting the organization of the cellular networks is present. The lack of an organization and of a user-friendly presentation of data complicates and limits their utilization. To overcome these weaknesses, we designed and implemented a cellular network relational database for fingerprint positioning systems, which integrates a large set of relevant data about cellular network, at different levels of granularity, in a coherent and systematic way.

It is well known that signal fingerprinting heavily relies on a comprehensive and accurate knowledge of cellular network configurations. In view of that, we start with an in-depth analysis of cellular networks in order to define a conceptual schema able to capture all meaningful aspects of their organization (Sect. 2). Then, we develop (Sect. 3) and populate (Sect. 4) the database, and show, by means of some representative examples, how useful information about the configuration of the network can be easily obtained from it (Sect. 5).

## 2 Basics of Cellular Networks

Cellular networks support wireless communication between mobile devices (both voice and data transmissions), and allow for seamless nation or even worldwide roaming with the same mobile phone. Different cellular technologies have been proposed over the years, each one with its own distinctive features [6, 9, 11].

Cellular radio networks are based on the deployment of a large number of low-powered base stations for signal transmission, each one with a limited transmission area, covering the surroundings with typically more than one cell. Cells are grouped into clusters to avoid adjacent cells to use the same frequency. Usually, a cell overlaps one or more other ones; a mobile device can distinguish among them by making use of their frequencies and scrambling codes (in the case of UMTS and LTE). Cells in a mobile network are put together into administrative areas, known as Location Areas (LA) in 2G/3G voice services, Routing Areas (RA) in 2G/3G data services, and Tracking Areas (TA) in 4G networks. These administrative areas are used to determine in a rough way the current location of a mobile device in the idle mode, that is, when it is switched on, but it is not using the network for any call or data exchange.

Independently of the adopted technology, there is a Public Land Mobile Network (PLMN), which can be identified by the Mobile Country Code (MCC),

which indicates the country where the network is located, and the Mobile Network Code (MNC), which identifies the network in the country. On the basis of the administrative organisation, a Global Cell Identifier (GCI) can be used to globally identify every cell. Even though the properties of each element depend on the specific technology, the GCI can be viewed as the concatenation of the PLMN identifier, the LAC or the TAC identifier, and the Cell Identifier (CI).

### 3 A Multilevel Database for Cellular Networks

Following the consolidated methodology for database design, we start with the conceptual schema, which has been developed by using the ChronoGeoGraph (CGG) model [3, 4], a spatiotemporal extension of the Entity-Relationship model.

**Conceptual Design.** The analysis of cellular networks reported in Sect. 2 makes it clear that they are organized in a hierarchical way (Fig. 1 - left side). Each network (entity *PLMN*) can be univocally identified by the values of the attributes *mcc* and *mmc*, and it consists of a number of distinct components (“subnetwork” entity *subPLMN*). Such a decomposition depends on the specific cellular technologies. According to the administrative perspective, each subnetwork consists of a number of cells (entity *CELL*), grouped into administrative areas (entity *ADMINISTRATIVE AREA*), on the basis of their registration and/or routing services. The properties of each level of the network organization depend on the specific technology. The differences among technologies are modeled by means of a suitable specialization of the entity *subPLMN*. The most significant one occurs at the level of the administrative areas (Fig. 1 - right side): the first two generations (2G GSM and 3G UMTS) distinguish two kinds of area, namely, Location and Routing Areas (entities *LA* and *RA*, respectively), which are identified by specific *lac* and *rac* codes. In the fourth generation, Location and Routing Areas are replaced by the Tracking Area (entity *TA*), which is characterized by a *tac* code. In all cases, a single code, combined with the *ci*, is used to identify the cells: the *lac*, for 2G and 3G, and the *tac*, for 4G.

**Logical Design and Implementation.** The CGG schema is turned into a relational one by applying the standard rules for the ER-to-relational schema mapping paired with dedicated rules for the encoding of the CGG spatial features [4]. The resulting schema is then implemented in the DBMS PostgreSQL with its spatial extension PostGIS. PL/SQL triggers are used for the automatic population of the tables. More precisely, when a new observation is received, it is checked and, if valid, the corresponding cell is inserted (or updated) in the table of cells. When a cell is inserted (or updated), the corresponding administrative area is inserted (or updated) in the table of the location/tracking areas, and so on. To obtain a comprehensive description of cellular networks via observations, the logical schema has been extended with some derived attributes whose value can be automatically computed, such as timestamp attributes (*firstview* and *lastview* attributes), that model the lifespan of the instances, as inferred from

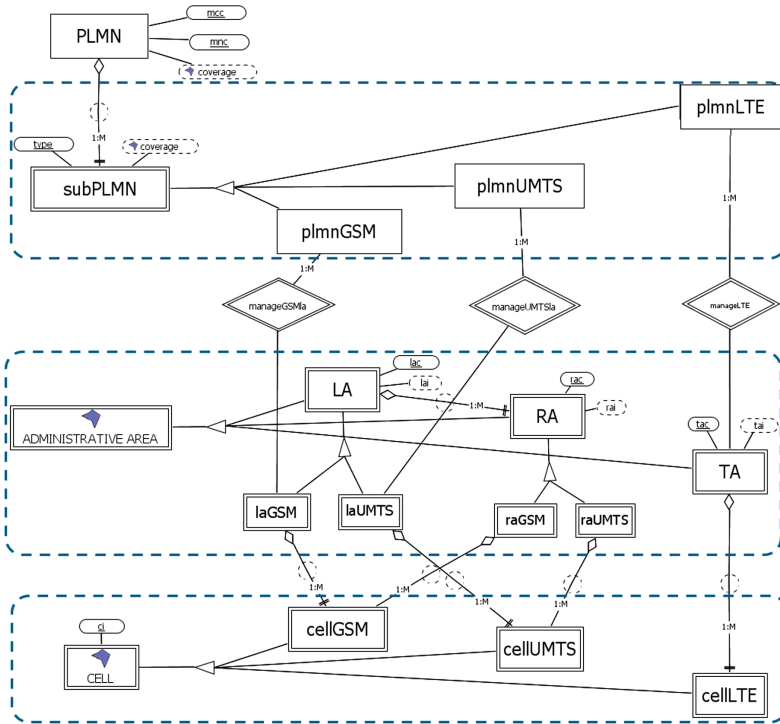


Fig. 1. A hierarchical schema for cellular networks: the administrative organization.

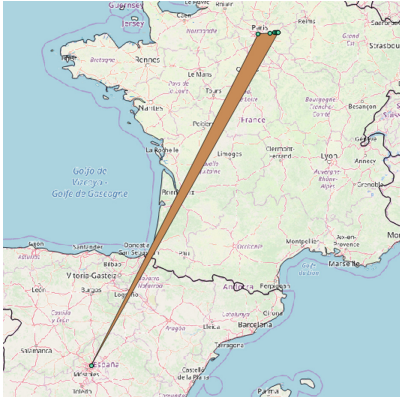
observations, and a counter attribute (attribute `numobs`), which keeps track of the number of observations of a single component to evaluate its reliability.

### 4 Data Ingestion and Filtering

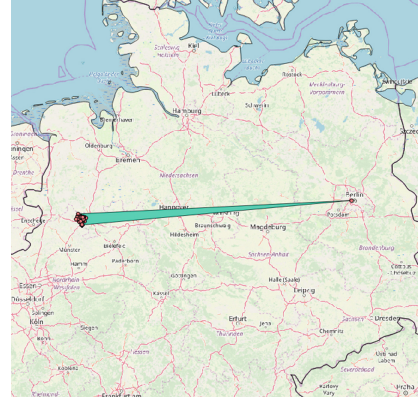
In order to test our proposal, we used the OpenCellID dataset [13], which is probably the most popular, publicly available, crowd source project in the field. Basically, users collect observations about signals from cell towers and locations using a specific mobile phone application. On 2017, the project was acquired by the Unwired Labs company, a geolocation service provider enterprise that affected both the privacy policies and the level of detail of published data.

We downloaded the OpenCellID dataset on April 2017, just before the change to the data and privacy policies. Data is organized in a tabular format (`csv` format), where each observation is encoded by means of a number of attributes: `mcc`, `net`, `area`, `cell`, `lon`, `lat`, `signal`, `measured`, `created`, `rating`, `speed`, `direction`, `radio`, `ta`, `rnc`, `cid`, `psc`, `tac`, `pci`, `sid`, `nid`, and `bid`. As a matter of fact, not all these attributes are available for all technologies, and even their meaning slightly changes from one technology to the other. Moreover, since several devices contributed to the dataset, there are significant differences in terms of attribute

structure and attribute subsets. The dataset covers an interval of about three years (from 2014-01-01 to 2017-03-17). It includes 42,951,377 observations distributed among three different cellular technologies. On the basis of a preliminary cleaning step, we restricted ourselves to 26,840,87 GSM observations, 6,177,024 UMTS, and 9,848,455 LTE (total 42,865,566) over the entire globe.



**Fig. 2.** Frequent errors: oversized cell over country.



**Fig. 3.** Frequent errors: oversized cell within country.

Data from a real scenario is generally affected by different types of error. As noticed in [12], there are at least four phenomena leading to incorrect readings: erroneous Cell IDs, antenna dragging, outliers, and unrealistic cell sizes. As a consequence, one of the most important operation in data acquisition is cleaning. To this end, we developed two filters to detect two relevant error situations. The first one deals with erroneous Cell IDs, and it consists of a domain check for each cellular parameter. The study of the cellular network standards allowed us to list a complete set of domains for each technology, e.g., `ci` ranges from 0 to 65535 for 2G and 3G, and from 65535 to 268435455 for 4G. If an observation exhibits one or more attribute values out of range, this is a valid reason to discard the entire observation, as some error may have occurred. The second filter focuses on the quality of the GPS position associated with each observation. First, to avoid inaccurate GPS positions, we excluded observations with less than 3 visible satellites. Later, a more interesting check has been done by comparing the GPS location of observations with borders of the country corresponding to the associated `mcc`. This spatial filter allowed us to avoid errors as the outlier depicted in Fig. 2. It is clear that a single cell belonging to a country cannot be received from such a long distance, in the middle of another country.

One of the main advantages of the proposed database is that, by keeping the geometry of cells constantly updated, it allows one to easily integrate additional filters. As an example, we may think of a filter that excludes observations which

are not coherent with the physical characteristics of the cellular network. Consider the scenario in Fig. 3, where the coverage of the shown cell is clearly too wide with respect to the transmitting range. This can be interpreted as an error, which may be caused by various phenomena (e.g., a device may have submitted discrepant positions and signals due to a failure or a switch on after a flight).

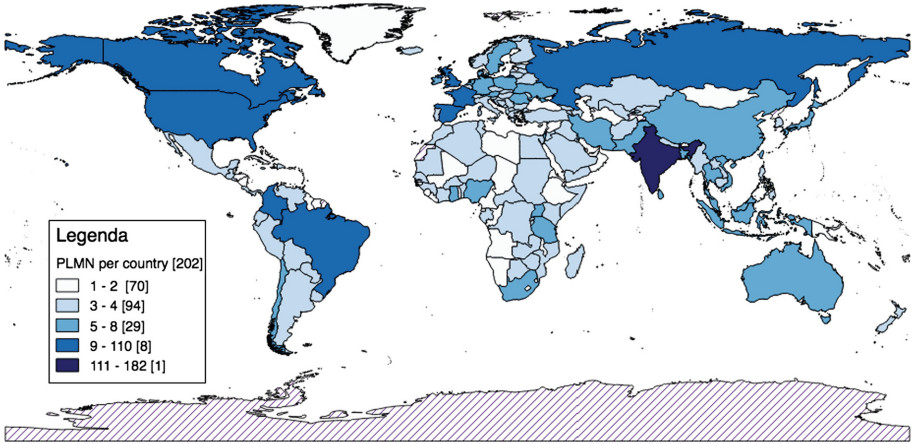


Fig. 4. Cellular networks per country.

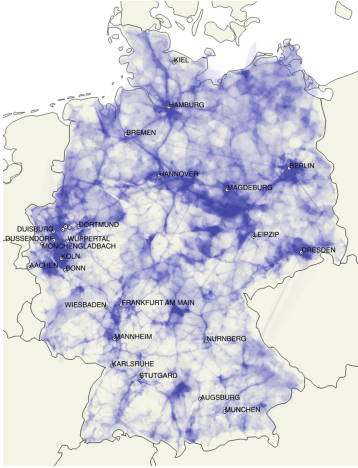
## 5 Data Analysis

Once the data ingestion and filtering phases have been completed, the hierarchical structure of the schema allows us to execute some meaningful analysis tasks on the filtered data. It is worth pointing out that all the analyses rely upon a reconstruction of the actual network as perceived through the observations recorded in the OpenCellID dataset. Moreover, despite the application of a couple of filters, that ruled out some inconsistent data, most probably data are still affected by errors. However, our main goal is to validate the proposed data model, and the above limitations have a little impact on it.

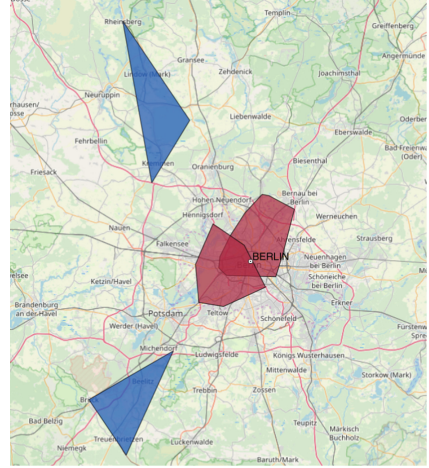
Let us first focus on the PLMNs occurring at the coarsest level. In Fig. 4, we give a graphical account of the result of a query that computes the number of cellular networks available in any single country, ignoring the specific cellular technology. India turns out to be the country with the maximum number of PLMNs (182), followed by USA (110) and Brazil (17). If we consider density (number of PLMNs divided by the area of the country), the first positions are occupied by small countries like Monaco, Gibraltar, and Macau.

Let us consider now the finer levels of the schema where administrative areas and cells come into play. In [10], the authors make it evident the existence of a correlation between the size of administrative areas and the density of population

in the area. Other experiments showed that the number of cells in a given administrative area can also be used to distinguish between rural and urban areas. This is the case, for instance, with the results reported in [16], where the density of base stations, and thus of cells, is taken as a good criterion to classify an area as rural or not. By exploiting the relationships between the various spatial entities, we can easily determine the internal composition of the administrative areas, and compute the density of the cells.



**Fig. 5.** Density of cells.



**Fig. 6.** Urban (red) and rural (blue) areas. (Color figure online)

In Fig. 5, we give a graphical account of the density of cells, in relation to the localization of urban areas, as it emerges from the pieces of information recorded in the considered dataset. It is clear from the picture that areas with a high density of cells (dark purple) are located where urban areas are, in particular where the cities with the highest population, such as Berlin, Hamburg, and the area of Köln, are (the largest cities are labeled with their name). As another example, working with GSM location areas with an extension between 50 Km<sup>2</sup> and 500 m<sup>2</sup> (avoiding areas which are not significant for the lack of a sufficient number of observations or for the presence of errors), it is possible to select the two areas with the highest (reps. lowest) density. As already pointed out, we cannot assume the dataset to be complete, and thus we may expect new cells to be added in the areas under consideration. However, the overall result confirms the original idea, as shown in Fig. 6: the two areas with the highest density (red areas) are located in the urban area of Berlin; the two areas with the lowest density (blue areas) are located in the rural area near Berlin.

## 6 Conclusions and Further Work

Thanks to their global coverage, cellular networks play a relevant role in a number of different contexts, including positioning systems based on cellular fingerprint observations. A comprehensive and accurate knowledge of their configuration is thus extremely important to optimise their usage. This paper is a first step towards the creation of a data store providing carefully structured information coming from observations of cellular networks.

We focused on the administrative organization of the networks (pieces of information usually available on the mobile device side), and we proposed a multilevel database that can be automatically updated with new cellular fingerprint measurements. We implemented the database in PostgreSQL, taking advantage of its spatial extension PostGIS, and populated it with an open data collection. Then, we demonstrated by some representative examples that it allows one to filter out inconsistent data and to perform a number of meaningful analyses.

We are currently exploring various possible improvements to the work done. One is the integration of other (public and proprietary) data sources in the database. We are also investigating an extension of the database schema with information about the network architecture as well as the physical parameters. Finally, to systematically deal with changes of the network configuration over time [14], we are thinking of adding some temporal dimensions [3, 5].

## References

1. Benikovsky, J., Brida, P., Machaj, J.: Localization in real GSM network with fingerprinting utilization. In: Chatzimisios, P., Verikoukis, C., Santamaría, I., Ladomada, M., Hoffmann, O. (eds.) *Mobilight 2010*. LNCS, vol. 45, pp. 699–709. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-16644-0\\_60](https://doi.org/10.1007/978-3-642-16644-0_60)
2. Chen, M.Y., et al.: Practical metropolitan-scale positioning for GSM phones. In: Dourish, P., Friday, A. (eds.) *UbiComp 2006*. LNCS, vol. 4206, pp. 225–242. Springer, Heidelberg (2006). [https://doi.org/10.1007/11853565\\_14](https://doi.org/10.1007/11853565_14)
3. Gubiani, D., Montanari, A.: ChronoGeoGraph: an expressive spatio-temporal conceptual model. In: *Proceedings of the 15th SEBD*, pp. 160–171 (2007)
4. Gubiani, D., Montanari, A.: A tool for the visual synthesis and the logical translation of spatio-temporal conceptual schemas. In: *Proceedings of the 15th SEBD*, pp. 495–498 (2007)
5. Gubiani, D., Montanari, A.: A relational encoding of a conceptual model with multiple temporal dimensions. In: Bhowmick, S.S., Küng, J., Wagner, R. (eds.) *DEXA 2009*. LNCS, vol. 5690, pp. 792–806. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-03573-9\\_67](https://doi.org/10.1007/978-3-642-03573-9_67)
6. Hoy, J.: *Forensic Radio Survey for Cell Site Analysis*. Wiley, New York (2013)
7. Paek, J., Kim, K.-H., Singh, J.P., Govindan, R.: Energy-efficient positioning for smartphones using cell-id sequence matching. In: *Proceedings of the 9th MobiSys*, pp. 293–306 (2011)
8. Li, X., Zhang, X., Chen, K., Feng, S.: Measurement and analysis of energy consumption on android smartphones. In: *Proceedings of the 4th ICIST*, pp. 242–245 (2014)



9. Pahlavan, K., Krishnaumrthy, P.: Principles of Wireless Access and Localization. Wiley, New York (2013)
10. Ricciato, F., Widhalm, P., Craglia, M., Pantisano, F.: Estimating population density distribution from network-based mobile phone data (2015)
11. Sauter, M.: From GSM to LTE: An Introduction to Mobile Networks and Mobile-Broadband. Wiley, New York (2011)
12. Ulm, M., Widhalm, P., Brändle, N.: Characterization of mobile phone localization errors with OpenCellID data. In: Proceedings of the 4th ICALT, pp. 100–104 (2015)
13. Unwired Labs: OpenCell ID (2017). <http://www.opencellid.org>. Accessed 28 Feb 2019
14. Viel, A., et al.: Dealing with network changes in cellular fingerprint positioning systems. In: Proceedings of the 7th ICL-GNSS, pp. 1–6 (2017)
15. Zhuang, Z., Kim, K.-H., Singh, J.P.: Improving energy efficiency of location sensing on smartphones. In: Proceedings of the 8th MobiSys, pp. 315–330 (2010)
16. Zhou, Y., et al.: Large-scale spatial distribution identification of base stations in cellular networks. *IEEE Access* **3**, 2987–2999 (2015)