Mobile Data as Public Health Decision Enabler: A Case Study of Cardiac and Neurological Emergencies

Sadok Ben Yahia[a], Gayo Diallo[b], M. Pathé Diallo[b], Ross Gore[c], Jyri Hämäläinen[d], Vianney Jouhet[b], Chiheb Karray[a], Nouha Kheder[a], Fleur Mougin[b], Edward Mutafungwa[d], Rym Saddem[a], Frantz Thiessard[b]

[a] Faculté des Sciences de Tunis University of Tunis, Tunisa
[b] ERIAS INSERM U897, University of Bordeaux, F-33000, France
[c] Virginia Modeling Analysis and Simulation, Old Dominion University, VA, USA
[d] Department of Communications and Networks, Aalto University School of Electrical Engineering, Espoo, Finland

Abstract

The establishment of hospitals in an area depends on many parameters taken into account by health authorities. We would like to investigate whether data from the use of mobile phones could feed this reflection. In order to do this, we chose two diseases that require rapid hospitalization for their care: myocardial infarction and stroke. The objective of the study is to show the areas in which the absence of a nearest hospital can result in death or serious sequelae.

In the approach we propose, the antenna coverage was estimated by the use of Voronoi diagrams. The real population density in each antenna area was estimated with the mobile population density. A total of 40 hospitals located across the 14 regions of Senegal where considered for the study. The maximum distance around each hospital was estimated to be reached in 90 minutes or three hours (corresponding to the time limit for the two diseases considered). The numbers of expected cases for the two diseases were estimated with the incidence rates of stroke and Myocardial infarction in the population, and the number of people in each antenna area. As a result, from the expected 13508 strokes each year, only 462 (3.42%) will occur too far from a hospital to be able to have the trombolysis treatment, because 96% of the population can reach a hospital in less than 3 hours. By cons, from the expected 24315 Myocardial infarctions, 4241 (17.4%) will occur too far from a hospital to be able to have the baloon treatment because they can't reach the hospital in less than 90 minutes.

Keywords: *Mobile data, Public Health, Stroke, Myocardial Infarction, D4D Challenge Senegal*

I. Background

Some medical emergency require rapid hospitalization for their care. We chose two diseases that fall within this framework and with a known maximum time limit for the treatment: myocardial infarction and stroke.

Myocardial infarction is an absolute cardiological emergency, the incidence remains high with 120 000 cases per year in France. According to WHO data, ischemic heart disease is the leading cause of death with 7.2 million of coronary heart disease death over the 50 million annual deaths worldwide. The prognosis remains serious since the IDM is still responsible for 10 to 12% of total annual adult mortality. In case of myocardial infarction, it is possible to perform a mechanical unblocking by the expansion of a balloon in a coronary to be carried out within 90 minutes after the first signs of the crisis.

Stroke is the leading cause in Western countries of acquired disability in adults, the second cause of dementia after Alzheimer's disease (30% of dementias are wholly or partly due to stroke), and the third leading cause mortality. In Europe, the annual incidence of stroke is between 101 and 239 per 100 000 for men and between 63 and 159 per 100 000 for women. In developing country, such as Senegal, the burden of stroke and other non-communicable diseases has risen sharply. In Dakar, stroke is the most frequent neurological disease with the highest mortality.

Takling health related issues, thanks to the recent advancements in data analysis over huge amount of data sources, almost all the determinants of our health – from our individual genetic coding to our particular habits – is becoming knowable. In that context, Big data may be the future for healthcare. Besides the possibility of achieving personalized medicine (FDA, 2013), cross-linking and analyszing various heterogeneous data sources can help early identification of factors that influence people health.

In this paper, we propose an approach based on the use of huge amount of recorded and anonymized mobile data to identify and estimate population at risk of major Public Health issues in particular stroke and myocardial infarction. To that end, we rely on data provided in the context of the Data For Development (D4D) challenge launched in 2014 by Orange France Telecom[1]. This year is the second edition and Senegal is the country concerned.

Senegal is a West African country bordered by 5 countries (Gambia, Guinea-Bissao, Guinea, Mauritania and Mali) and totalizing 196,712 km^2. The country is subdivised in 14 regions. It is further subdivided by 45 Départements, 123 Arrondissements and by Collectivités Locales. The total population is estimated to 13,508,715 people according to the last General Census of the Population. The capital city is Dakar. It concentrates the main busness activity of the country. We provide in Table 1 main figures about the population in Senegal (RGPHAE, 2013).

II. Materials

Dataset used

[1] http://d4d.orange.com/en/home

Orange Senegal Mobile Data

The dataset provided by Orange Senegal are based on fully anonymized Call Detail Records (CDR) of mobile phone calls and SMS between the company clients in Senegal between January 1st 2013 and Decembre 31st 2014 (• Montjoye et al. 2014). The collected CDR which initially comprises 9 million unique aliased mobile phone numbers have been reduced following two criteria (Montjoye et al. 2014):

- users having more than 75% days with interactions per given period (biweekly for the second dataset, yearly for the third dataset)
- users having had an average of less than 1000 interactions per week. The users with more than 1000 interactions per week were presumed to be machines or shared phones.

Dataset 1: it contains metadata about the traffic between each antenna for 2013. It includes both voice and text traffic. Table 2 and y gives examples for voice traffic between sites traffic.

Table 2: Example of voice traffic between antennas

Timestamp	Out-going site	Inco-ming site	Number of calls	Total call du-ration
2013-04-01 00	2	2	7	138
2013-04-01 00	2	3	4	136
2013-04-01 00	2	4	7	121
2013-04-01 00	2	5	13	272
2013-04-30 23	1651	1632	1	3601
2013-04-30 23	1653	575	1	20
2013-04-30 23	1653	1653	2	385
2013-04-30 23	1659	608	1	3601

Table 1: Main figures about population in Senegal

Region Number on the map	Name of the region	Number of males	Number of females	Global Popu-lation	Area (km²)	Density (/km²)
1	Dakar	1 579 020	1 558 176	3 137 196	547	5735.3
2	Thies	896 572	892 292	1 788 864	6670	268.2
3	Diourbel	716 460	780 995	1 497 455	4824	310.4
4	Kaolack	474 404	486 471	960 875	5357	179.4
5	Saint-Louis	453 315	455 627	908 942	19241	47.2
6	Louga	433 715	440 478	874 193	24889	35.1
7	Fatick	353 716	360 676	714 392	6849	104.3
8	Tambacounda	344 475	336 835	681 310	42364	16.1
9	Kolda	335 018	327 437	662 455	13771	48.1
10	Kaffrine	282 093	284 899	566 992	11262	50.3
11	Matam	276 481	286 058	562 539	29445	19.1
12	Ziguinchor	281 813	267 338	549 151	7352	74.7
13	Sedhiou	229 468	223 526	452 994	7341	61.7
14	Kedougou	78 867	72 490	151 357	16800	9.0
	Total	6 735 417	6 773 298	13 508 715	196 712	68.7

Dataset 2: this dataset contains two weeks basis fine-grained mobility data. It is constituted of the trajectories at site (antenna) level of about 300,000 randomly selected users meeting the two previously mentioned criteria. Table 3: example of dataset 2 gives an example for a given user of the dataset which comprises 25 different files.

Orange provides also a coarse-grained mobility dataset (referred to as dataset 3) which contains trajectories at arrondissement level. We have not used it in the current study.

Table 3: example of dataset 2

user	timestamp	site
1	18/03/2013 21:30	716
1	18/03/2013 21:40	718
1	19/03/2013 20:40	716
1	19/03/2013 20:40	716
1	19/03/2013 20:40	716
1	19/03/2013 20:40	716
1	19/03/2013 21:00	716
1	19/03/2013 21:30	718
1	20/03/2013 09:10	705
1	21/03/2013 13:00	705

Contextual Data

In addition to CDR related data, the administrative organization of Senegal is provided as well as the antenna GPS coordinates and the arrondissement to which they belong to.

We have also used data from the National Agency of Statistics and Demographics, in particular the last available Senegal General Population and Housing Census (RGPHAE 2013).

The list of hospital is compiled from the online Senegal Medical Directory (SMD, 2014) and the SenDoctor web site (SD, 2014).

Eventually, in addition to the shapefiles for Senegal provided by the D4D challenge organizers, we used data from OpenStreetMap.org (OSM, 2014).

III. Description of the approach

The overall followed methodology in our study is conducted in regards with the following hypothesis:

- The average incidence rate of stroke in Senegal is estimated to 100 per 100 000 inhabitants. As there is no recent documented study which gives us figures that could be used, we based our hypothesis on the fact that the incidence of stroke in Europe is between 63 and 159 for 100 000 women, and between 101 and 239 for 100 000 men. For Senegal it is supposed to be less.
- The average incidence rate of myocardial infarction is estimated to 180 per 100 000 inhabitants in France. Based on that, we assume that the incidence rate in Senegal is about 150 per 100 000 inhabitants
- We base our estimation distance from home to the nearest hospital in case of emergency of stroke on the French High Authority for Health (HAS) recommendations (HAS, 2014). The recommendations estimate that for a severe stroke, the patient needs to taken in charge no more than 3h.
- For myocardial infarction the maxim time for an efficient management is estimated to 1h30 (90 minutes).

Estimating antenna coverage

The geographical coverage areas of mobile (cellular) networks have often been described using equal-sized hexagonal coverage areas around each cell site. These hexagonal grid models have routinelly been used for system performance studies for instance in Third Generation Partnership Project (3GPP) standardization studies (Holma and Toskala 2009). However, in reality the cellular layout is highly irregular due to constraints on the where the cell site could be located, spatio-temporal variations in mobile penetration and population density, the surrounding topography, presence of buildings, and so on (Holma and Toskala 2009).

The use of Voronoi diagrams have been proposed as tesselation that overcomes the inaccurate hexagonal grid cellular representation when compared to real world cellular network layouts (Baert and Seme 2004). In the Voronoi diagram approach, cellular network for an area covered by N cells sites area is subdivided into convex polygonial regions around N points that correspond to the locations of the N sites. The irregular shape of the polygons allows provides a relatively better approximation of cell size by taking into account irregular site locations and promity of neighbouring sites. The Voronoi tessalation generated for the provided 1666 cell sites in Senegal is shown in Figure 1 below.

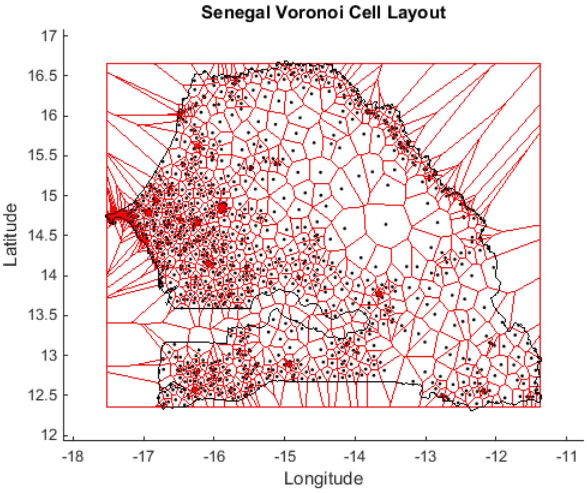

Figure 1 Voronoi cell layout for Senegal based on provided 1666 site locations

Computing mobile population density

In order to have an overall view of the distribution of mobile population at regional level, we use data from dataset 2. The idea is to agregate a daily average mobile phones identified in a particular antenna. We make the hypothesis that census corresponds to the place where people are globally. We use a first correction factor α for adjusting identified unique users to the 300.000 2 weeks based Orange users.

We then use a first correction factor β to ajust the number of people phoning to the expected number of people in the area according to the census provided by (RGPHAE 2013) and the Orange market share in Senegal.

$$\alpha = U_S * 1.2 * {}^{1}/_{Oms}$$

$$\beta = \frac{U_R}{O_R}$$

Where

- U_S is the number of unique users per day and per antenna computed from the dataset 2
- Oms is the estimated orange market share in 2013 according to GSMA survey[2].
- The factor 1.2 is obtained by adjusting the totatl numbers of unique users to 300.000 as of dataset 2.

The α and β adjustment coefficients are important to make the corrections of all counted number of people calling during a given day to have an idea of the real number of people at each place (antenna location) (and the total number of people should be 13 508 715 according to population census). For instance, the β correction factor for the Dakar region is 9.05 while it is 160.11 for the Sedhiou region and even 509.43 for the Kedougou region. Table 4: Average number of unique users per site daily gives an example of estimated population by antenna site.

Table 4: Average number of unique users per site daily

Site	Region	Unique User/Day
1	Dakar	160,24
1583	Tambacounda	170,05
1405	Saint-Louis	181,04

[2] http://www.gsma.com/

Figure 2 represents distribution of antenna as well as the density of the population across the country. Lets assume that N represente number of people in a given antenna coverage. We have used 5 different colors respectively grey N < 100, yellow for 101<N<1000, red for 1001<N<10000, brown for 10001<N<100000 and black for N > 100000.

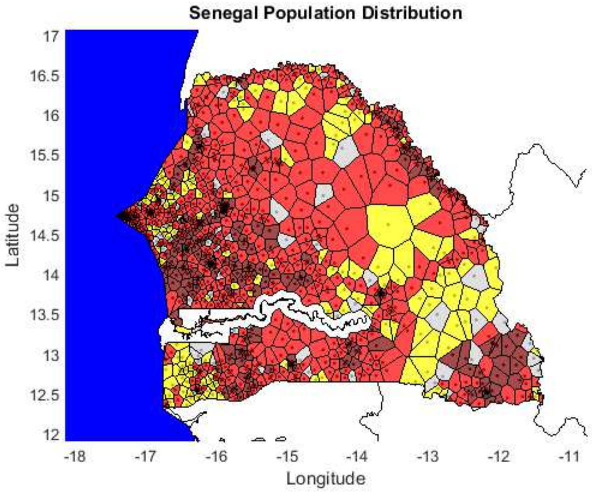

Figure 2: Distribution of Senegal population according to the antennas

IV. Estimating risk zones

Many health geographers use distance as a simple measure of accessibility, risk, or disparity in terms availability of health services in different locations (Dummer 2008). Time-critical medical emergencies like heart attacks and stokes considered in this study require that hospitals that provide emergency care are within a certain distance of the victims requiring immediate care. For the case of medical emergencies due to heart attacks and strokes Locations from which victims are unable to reach the hospitals within a given time are considered to be high risk areas.

In this study we utilize the mobile datasets provided to evaluate the areas are considered high risk based on given time criteria. This involes mapping the hospital and populated distribution on to the cellular network layout. A total of 40 hospitals located across the 14 regions of Senegal where considered for the study (see Figure 3 with the capital city Dakar highlighted). These hospitals have been retrieved from the online Senegal Directory and the SenDoctor web site. The geographical location of the hospitals was approximated by representing them using site IDs of the nearest antenna site. Using this approximation it was noted that 85% of the 40 hospitals considered where within 2 km of their real geographical locations (see Figure 4). The impact of this error is minimal when evaluating the travel time to the hospitals.

The segmentation of locations according to their proximity to hospitals is also done based on cell areas. To that end, a common travel time is assumed reach a hospital from a particular for all people located in the same cell area. Furthermore, the antenna site location is assumed to the centroid of the cell area and the distance to the hospital for cell is the distance between the cell antenna site and the antenna site associated to with the hospital.

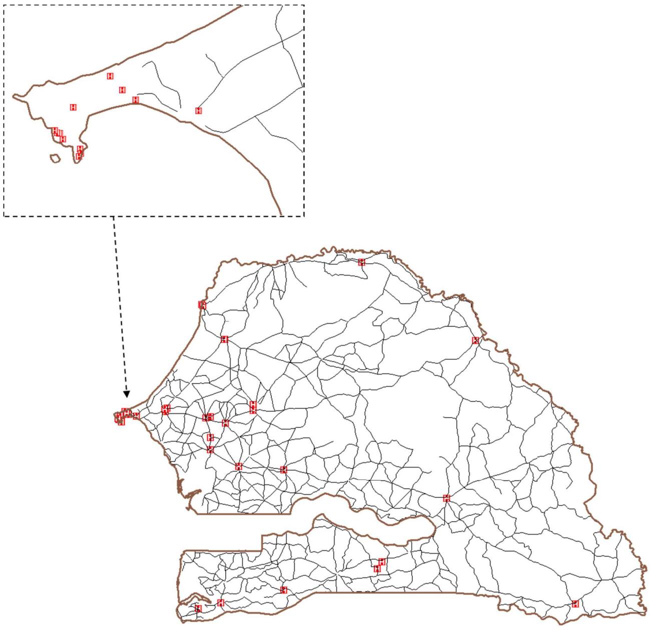

Figure 3 Location of the 40 hospitals in considered in the study. The hospitals are represented in red symbols while the road network is shown in black lines. The Dakar region is shown inset.

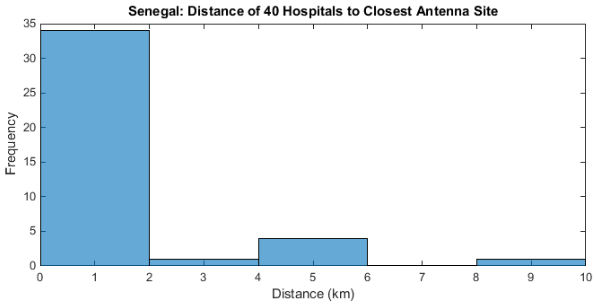

Figure 4 Distance of hospitals to closest antenna sites

A significant number past health planning studies have used used the straight-line ("as the crow flies") the measure the distance between two ponts on the map (Boscoe et al 2013). The approach uses either the spherical distance for geo coordinates (latitude and longitude) or Euclidean distance for projected coordinates. However, the real drive distances (and hence travel times) between the two points tend be longer due to the fact that roads are built around natural obstacles, such as, mountains, boulders and so on. This is clearly visible in the road network of Figure 3. To that end, a correction factor know as *detour index* representing the ratio of the drive-distance to the to the straight-line distance has been introduced to obtain more accurate distance estimates with less computation or measurement effort (Boscoe et al 2013). The detour index approaches the lower bound of 1 the denser the road network. In developed economies a detour indices in the range of 1.2 to 1.6 have been noted.

For this study we calculate the straight-line distance from each hospital to all cell sites and we then use a detour index of 2 to evaluate the drive distances between the points. The detour index assumption is rather conservative to take into account the relative low road network density and the less than ideal road conditions. Furthermore, for simplicity an average driving speed of 60 km/hr is assumed for all areas. An estimate of the travel time ranges from each cell area to the nearest hospital is illustrated in **Erreur ! Source du renvoi introuvable.**. The maximum of 90 and 180 minutes are considered based to the treatment time windows for the cardiovascular conditions considered in this study. The cell areas beyond the 180 minute catchment area are considered high risk areas for all conditions.

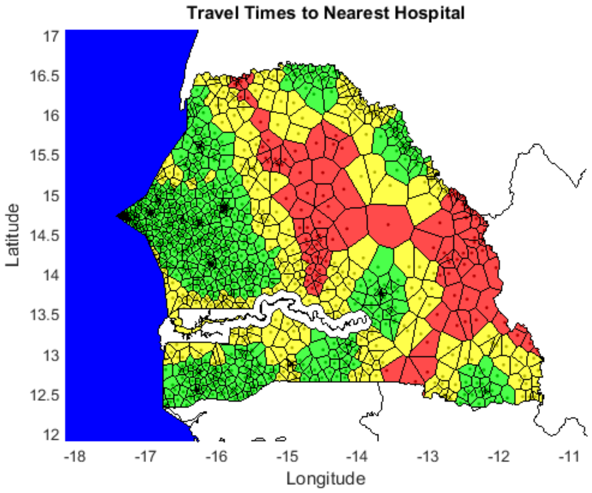

Figure 5 Estimated travel times from different areas to the nearest hospital. (Green: less than 90 minutes, Yellow: 90 to 180 minutes, Red: over 180 minutes).

From the expected 13508 strokes each year, only 462 (3.42%) will occur too far from an hospital to be able to have the thrombolysis treatment, because 96% of the population can reach an hospital in less than 3 hours (assuming that all the hospital are able to do the treatment) Figure 6.

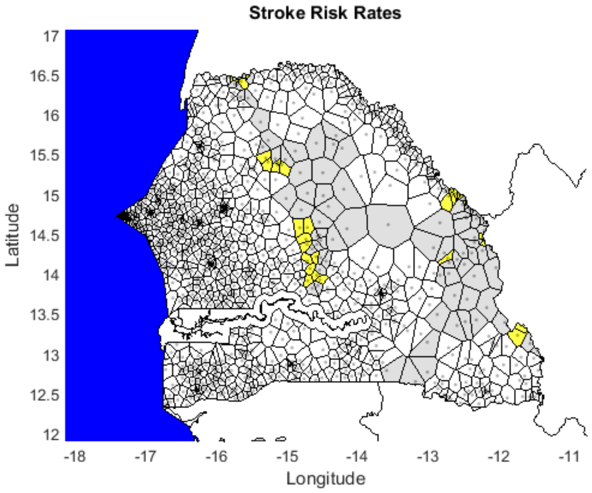

Figure 6. Estimated incident cases of strokes from different areas too far from the nearest hospital to be treated by balloon. (Grey < 5 victims, Yellow 5 to 25).

By cons, from the expected 24315 Myocardial infarctions, 4241 (17.4%) will occur too far from an hospital to be able to have the baloon treatment because they can't reach the hospital in less than 90 minutes (Figure 7).

V. Discussion

Highlights

Stroke and Myocardial Infarction are two majors issues in Public Health. Several factors contribute to the increasing of the number of the cases annually. This is particulalry true in countries such as Senegal.

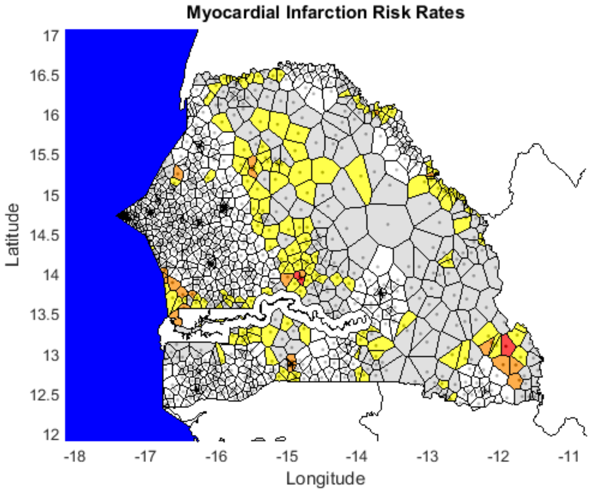

Figure 7. Estimated incident cases of Myocardial Infarction from different areas too far from the nearest hospital to be treated by fibrinolysis. (Grey < 5 victims, Yellow 5 to 25, Orange 26 to 50, Red 51 to 100).

We have shown that by using a considerable amount of recorded mobile data combined with census data it is possible to perform location based estimatimation of the people in risk. This will help Public Health decision makers in their early stage decision making.

Limitations of the study

The current study presents some limitations due to bias introduced by the data used and some choices for designing the study. The first bias is related to the extrapolation of population at a given antenna coverage are as it is not possible to estimate precisely the Orange share market in Senegal during the periode covered by the data. In addition, a filtering on data is performed by Orange as indicated in section 2. Another bias which may affect the results is related the computation unique users ber day for a given antenna site. We did not performed the estimation based solely on the users during night, which is likely to be more accurate. Eventually, we based the study on a estimated incidence rate of the considered medical emergency as there is no official figures.

Future work

There is a room for improvement of the current study. First, we envision to investigate a more fine-grained estimation of the population density. Indeed, currently we perform our estimation by using CDR mobility data at antenna level. Even if we ajust our figures with the census information, we do not take into account difference between night and day, neither more fine grained time frames according for instance to off pick hours.

For the medical emergencies considered in this study (stroke and myocardial infarctions) it has been assumed that all the considered hospitals possess the requisite capabilities for treatment of the emergencies. A more precise studies would take into account the individual treatment capabilities of each hospital to obtain more precise mapping of the high risk zones for the considered medical emergencies.

In order to generalize our approach to other emergency cases, we also plan to base our approach on a domain knowledge model represented by an ontology. The idea is to represent formally and semantically the different emergency case which may lead to death or irreversible sequelaes (cardiovascular diseases, stroke, etc.) and describe the different factors to take into account for an efficient management. To do so we will rely on semantic web technologies (Shadbolt et al., 2006).

Another issue that needs to be addressed is taking into account the ever growing available data through the Open Data initiative, which make available Linked Open Data (weather conditions, trafic jams, traffic networks, etc.). Coupled with the available Big mobile data, it should be possible to build on-demand or stream-based applications for takling major Public Health issues.

VI. Conclusion

In this paper, we have described our approach for the identification of risk zones for helping Public Health decion makers to take the required action on the earlier. Two major concerns in Public Health have been considered: myocardial infarction and stroke in the context of Senegal. Thanks to the use of anonymized mobile data provided in the context of the 2014 D4D challenge, we have been able to estimate population at risk.

VII. References

- Baert, A.-E. and Seme, D. 2004: Voronoi mobile cellular networks: topological properties, Third International Symposium on/Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks, 5-7 July 2004.
- Boscoe et al. 2013: A Nationwide Comparison of Driving Distance Versus Straight Line Distance to Hospitals, Professional Geographer, 64(2), 2013.
- Dummer, T. J. B. 2008: Health geography: supporting public health policy and planning. CMAJ : Canadian Medical Association Journal, 178(9), 2008.
- FDA, 2013. Paving the Way for Personalized Medicine: FDA’s Role in the New Era of Medical Product Development (October 2013)
- Holma H. and Toskala A. 2009: WCDMA for UMTS: HSPA Evolution and LTE, Wiley & Sons Ltd, Chichester, 2009.
- Montjoye et al. 2014 : D4D-Senegal : The Second Mobile Phone Data for Development Challenge. July 2014.
- OSM, 2014: Web Site OpenStreetMap, consulted on December 15, 2014.
- RGPHAE, 2013: Recensement Général de la Population et de l’Habitat
- SD, 2014 : SenDoctor http://sendocteur.com/structureregion.php, consulted on December 16, 2014
- SMD, 2014 : The online Senegal Medical Directory Web Site http://www.annuairemedical-senegal.com, consulted on December 30, 2014
- Shadbolt, N., Hall, W., Berners-Lee, L. "The Semantic Web Revisited", IEEE Intelligent Systems Journal, May/June 2006, 96-101

Acknowledgments

We would like to thanks the organizers of the D4D Challenge for providing us the data necessary to perform the current study. We also thanks medical doctors from Senegal who helped in getting information about health in this country.

Address for correspondence

Lead author contact details go here