



Sciences Economiques & Sociales de la Santé  
& Traitement de l'Information Médicale

[www.sesstim-orspaca.org](http://www.sesstim-orspaca.org)

**Frantz THIESSARD**  
**MD PHD Informatique médicale**  
**ERIAS - INSERM U1219 - ISPED - Université de Bordeaux**

**Exemple d'exploitation de données massives de  
téléphonie mobile au Sénégal pour une utilisation en santé**

janvier 2016



**Cliquez ici pour voir l'intégralité des ressources associées à ce document**

# Exemple d'exploitation de données massives de téléphonie mobile au Sénégal pour une utilisation en santé

Webinar QuanTIM

Frantz Thiessard & Gayo Diallo  
ERIAS ISPED INSERM U897

université  
de **BORDEAUX**

vendredi 22 janvier 2016

Lorsque les données proviennent de l'utilisation de téléphones mobiles ...

# Orange Data for Development Challenge in Senegal

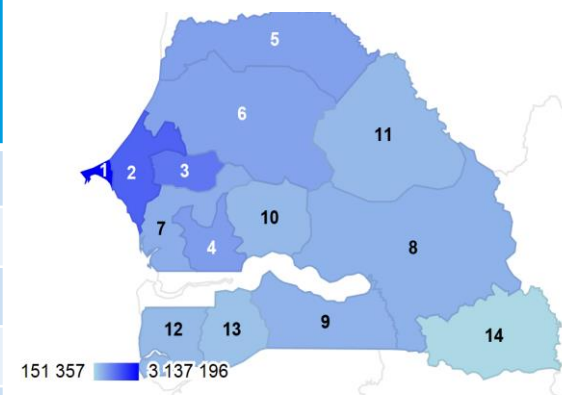


→ “Data for Development Senegal” est un défi ouvert pour l'innovation dans les TIC et le Big Data avec pour objectif le développement de la société.

Source: <http://www.orange.com>

# Sénégal: démographie et la découpage administratif

Numéro de la région sur la carte	Nom de la région	Effectif des hommes	Effectif des femmes	Population totale	Surface (km <sup>2</sup> )	Densité (/km <sup>2</sup> )
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
	<b>Total</b>	<b>6 735 417</b>	<b>6 773 298</b>	<b>13 508 715</b>	<b>196 712</b>	<b>68.7</b>



Source: General Population Census (RGPHAE) 2013

- Orange D4D Challenge
  - › Contexte
  - › Jeux de données et exemples de projets retenus
  
- Exploitation de données massives de téléphonie mobile au Sénégal pour des prises de décision en santé
  - › Objectif et contexte de l'étude
  - › Consortium and Données utilisées
  - › Approche et résultats
  
- Conclusion

# Orange D4D Challenge

- Lancé en 2012 par la compagnie Orange
- Partenariat avec l'Université de Louvain (Belgique) et le MIT (USA)
- Métadonnées de téléphonie mobile mises à disposition D4D :
  - Côte d'Ivoire (2012-2013) : 5 mois
  - Vs
  - Sénégal (2014-2015) : 12 mois
- ~360 projets soumis pour les deux challenges
- Articles produits :
  - 80 en 2013
  - Vs
  - 60 en 2015
- Divers sujets couverts
  - › optimisation des itinéraires de bus, en analysant les divisions sociales, étudier les politiques de confinement de maladies, etc.

# Détails sur le défi D4D Sénégal

## D4D-Senegal

- Sonatel et le Groupe Orange ont anonymisé les données disponibles, extraites du réseau mobile de l'opérateur au Sénégal
- De Avril 2014 à Avril 2015 (Soumission des projets : 31 Décembre 2014)
- 5 sujets prioritaires
  - › Santé
  - › Agriculture
  - › transports / urbanisme
  - › Énergie
  - › statistiques nationales

## D4D-Senegal: The Second Mobile Phone Data for Development Challenge

Yves-Alexandre de Montjoye<sup>1</sup>, Zbigniew Smoreda<sup>2</sup>,  
Romain Trinquart<sup>2</sup>, Cezary Ziemlicki<sup>2</sup>, Vincent D. Blondel<sup>3</sup>

<sup>1</sup>Media Lab, Massachusetts Institute of Technology, Cambridge, MA

<sup>2</sup>Orange Labs, France

<sup>3</sup>Université catholique de Louvain, Belgium

July 31, 2014

The D4D-Senegal challenge is an open innovation data challenge on anonymous call patterns of Orange's mobile phone users in Senegal. The goal of the challenge is to help address society development questions in novel ways by contributing to the socio-economic development and well-being of the Senegalese population. Participants to the challenge are given access to three mobile phone datasets. This paper describes the three datasets. The datasets are based on Call Detail Records (CDR) of phone calls and text exchanges between more than 9 million of Orange's customers in Senegal between January 1, 2013 to December 31, 2013. The datasets are: (1) antenna-to-antenna traffic for 1666 antennas on an hourly basis, (2) fine-grained mobility data on a rolling 2-week basis for a year with bandicoot behavioral indicators at individual level for about 300,000 randomly sampled users, (3) one year of coarse-grained mobility data at arrondissement level with bandicoot behavioral indicators

# Exemple de besoins dans le domaine des transports

- Identifier la demande de mobilité par région et par période de l'année,
- Mesurer la nécessité d'une infrastructure temporaire en raison d'événements ad hoc (match de football, pèlerinage)
- Optimiser l'infrastructure touristique
- Modèles des villes, et afficher les principaux courants d'optimisation
- Mesurer l'impact «avant-après» de l'introduction d'équipements. En particulier en ce qui concerne les événements suivants :
  - › Ouverture de l'autoroute Dakar en 2013
  - › Rénovation du réseau de chemin de fer





# Example of needs in the healthcare domain

- Cartographie des facteurs déterminants des maladies non transmissibles
- Accessibilité aux formations en santé
- Caractéristiques des besoins de soins de santé dans les milieux ruraux et urbains
- Corrélation entre pauvreté et santé
- Santé selon les groupes ethniques
- Corrélation entre accès à l'eau et santé
- Conditions de l'accès au soin (distance moyenne des centres de soin, emplacement des centres de vaccination, situation géographique des médecins, etc.)
- Simulation d'épidémies :
  - › Propagation du virus Ebola au Sénégal ? (Fermeture de la frontière avec la Guinée en raison de l'épidémie d'Ebola)

## → Critères d'inclusion

- › Année concernée : 2013
- › Plus gros utilisateurs du téléphone mobile avec plus de 75% des jours avec au moins un appel par période donnée
- › Utilisateurs ayant en moyenne moins de 1000 communications par semaine. (Hypothèse : utilisateurs avec plus de 1000 appels par semaine = machines ou téléphones partagés)

# Detail du jeu de données

## **Dataset 1**

Un an de communications de site à site pour 1666 sites sur une base horaire

## **Dataset 2**

Données fines de mobilité (niveau du site) par périodes de 2 semaines pour environ 300.000 utilisateurs tirés au sort

## **Dataset 3**

Un an de communication avec secteur plus grossier (123 arrondissement) données sur la mobilité de 150.000 utilisateurs tirés au sort

# Dataset 1: Trafic d'antenne à antenne, exemple

→ Trafic pour l'année: 25 fichiers compressés totalisant ~ 1 milliard de tuples

timestamp	outgoing_site_id	incoming_site_id	Nombre d'appels	Durée des appels
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

timestamp	outgoing_site_id	incoming_site_id	Nombre de sms
2013-04-01 00	2	12	16
2013-04-01 00	2	14	1
2013-04-01 00	2	21	1
2013-04-01 00	2	28	9

# Dataset 2: Mobilité fine

- 300.000 utilisateurs choisis au hasard, pour 25 périodes de deux semaines

<b>user_id,</b>	<b>timestamp,</b>	<b>site_id</b>
1,	2013-03-18 21:30:00,	716
1,	2013-03-18 21:40:00,	718
1,	2013-03-19 20:40:00,	716
1,	2013-03-19 20:40:00,	716
1,	2013-03-19 20:40:00,	716
1,	2013-03-19 20:40:00,	716
1,	2013-03-19 21:00:00,	716
1,	2013-03-19 21:30:00,	718
1,	2013-03-20 09:10:00,	705
1,	2013-03-21 13:00:00,	705

# Dataset 3: Mobilité « à gros grain »

- les trajectoires de déplacement (niveau arrondissement) de 146,352 utilisateurs tirés au sort

<b>user_id,</b>	<b>timestamp,</b>	<b>arrondissement_id</b>
37509,	2013-01-29 15:00:00,	3
84009,	2013-01-14 07:00:00,	3
84009,	2013-01-14 07:00:00,	3
84009,	2013-01-14 07:00:00,	3
80150,	2013-01-27 16:50:00,	3
52339,	2013-01-09 19:50:00,	48
52339,	2013-01-06 17:50:00,	48
52339,	2013-01-13 15:40:00,	48
52339,	2013-01-03 19:00:00,	48
52339,	2013-01-07 01:30:00,	48

# Données contextuelles

- Divisions administratives du Sénégal (fournies par l'ADSN)
- Données météorologiques
- Données démographiques et socio-économiques (RGPHAE 2013)

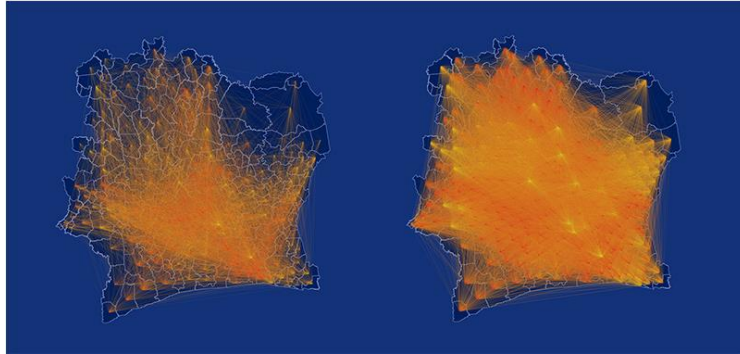
```
ARR_ID,REG,DEPT,ARR
1,DAKAR,DAKAR,PARCELLES ASSAINIES
2,DAKAR,DAKAR,ALMADIES
3,DAKAR,DAKAR,GRAND DAKAR
4,DAKAR,DAKAR,DAKAR PLATEAU
5,DAKAR,GUEDIAWAYE,GUEDIAWAYE
6,DAKAR,PIKINE,PIKINE DAGOUDANE
```

## Additional Sources

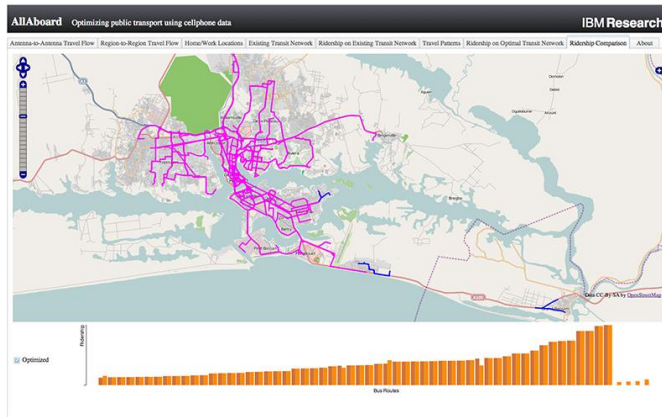
- National Statistics Agency
- Open Street Map
- Paris21
- Etc.

# Projets retenus en 2013

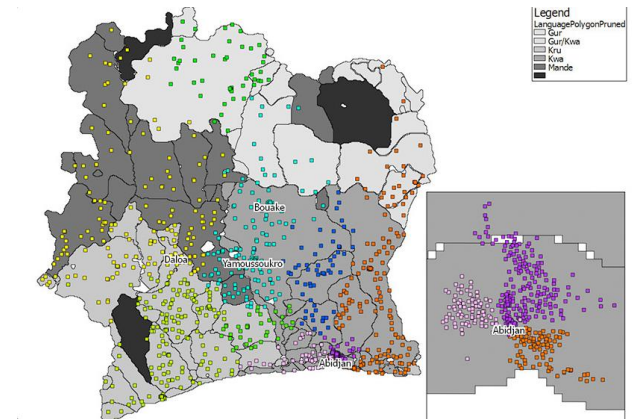
- 1: Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-Wide Epidemics [University of Birmingham]



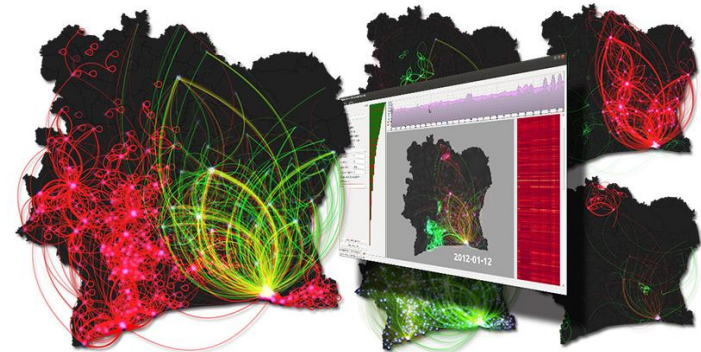
- 2: Tous à bord (AllAboard) : a system for exploring urban mobility and optimizing public transport using cellphone data [IBM, Dublin]



- 3: Analysis of social categories by using cell phones data [University of California/San Diego]



- 4: Study and analysis of massive mobile data [Eindhoven University of Technology/SynerScope BV/MIT]





# Some of the winners of 2015



## **First Prize and Energy Prize: Using mobile phone data for electrification planning**

E.A. Martínez-Ceseña <sup>(1)</sup>, P. Mancarella <sup>(1)</sup>, M. Ndiaye <sup>(2)</sup>, and M. Schläpfer <sup>(3)</sup>

Knowledge of local energy needs is crucial for the electricity infrastructure planning of a country. We have shown that mobile phone data are an accurate proxy of the energy needs and can be used to develop bottom-up demand models. The new methodology supports and prioritizes the electrification plans in areas with scarce information on local activities and energy consumption.

(1)University of Manchester, UK - (2) Ecole supérieure polytechnique de Dakar UCAD, Senegal - (3) Santa Fe Institute, USA



## **Agriculture Prize: Genesis of millet prices in Senegal: the role of production, markets and their failures**

D.C. Jacques <sup>(1)</sup>, R. d'Andrimont <sup>(1)</sup>, J. Radoux <sup>(1)</sup>, F. Waldner <sup>(1)</sup>, and E. Marinho <sup>(2)</sup>

Information asymmetries are responsible for price differentials in only the few areas where the mobile phone coverage has not yet reached its full potential, which damages both poor producers and food insecure consumers. To address this issue, we have integrated it in a spatially explicit model that simulates the functioning of agricultural markets.

(1) Earth and Life Institute, Université Catholique de Louvain, Belgium - (2) Independent researcher, Rio de Janeiro, Brazil



## **Transport Prize: National and Regional Road Network Optimization for Senegal Using Mobile Phone Data**

Y. Wang <sup>(1)</sup>, G. Homem de Almeida Correia <sup>(1)</sup>, and Erik de Romph <sup>(1,2)</sup>

Anonymous mobile phone traces can be filtered with an algorithm to generate a proxy for a trip origin-destination matrix. This is used to develop a gravity model that predicts the future mobility in the country dependent on travel time and number of calls and messages between the departments. This information is then used to improve decision making for road network planning.

(1) Department of Transport and Planning, Delft University of Technology, The Netherlands - (2) DAT.mobility, The Netherlands

# Large CDR Data as Public Health Decision Enabler: A Case Study of Cardiac and Neurological Emergencies



NetMob15 @netmob15 · Apr 10  
Groups awarded the D4D prizes 2/3

Practical Application Prize of D4D 2015

# Objectifs de l'étude

→ Le problème de l'emplacement optimal des Structures de soin pour maximiser la couverture de la population est crucial, en particulier pour les cas d'urgences médicales où le temps est critique

Nous évaluons les risques de complications liés à l'inaccessibilité aux soins dans un temps court pour des pathologies fréquentes

**Objectif de l'étude** : Montrer les zones géographiques dans lesquelles l'absence d'un hôpital peut entraîner la mort ou des séquelles graves, grâce à l'utilisation des données mobiles

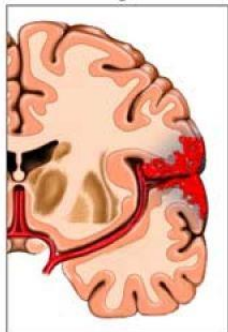
# Contexte de l'étude

- Certaines maladies nécessitent une intervention médicale très précoce
- Absence de soins => mort du patient ou séquelles graves
- 2 cas d'utilisation : AVC et infarctus du myocarde

**Stroke –**  
there's treatment if you act **FAST.**

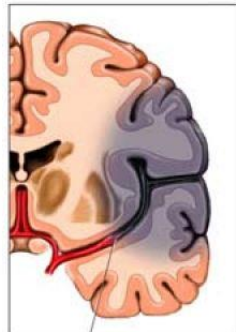


Hemorrhagic Stroke



Hemorrhage/blood leaks into brain tissue

Ischemic Stroke



Clot stops blood supply to an area of the brain



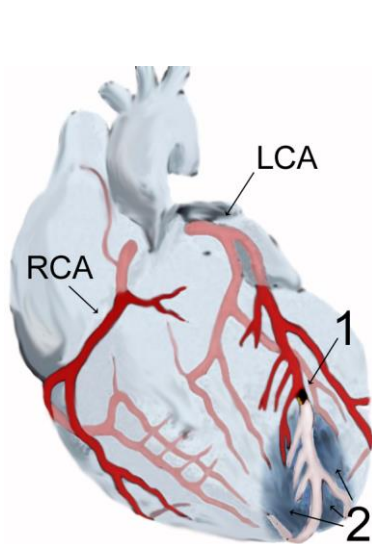
180 min



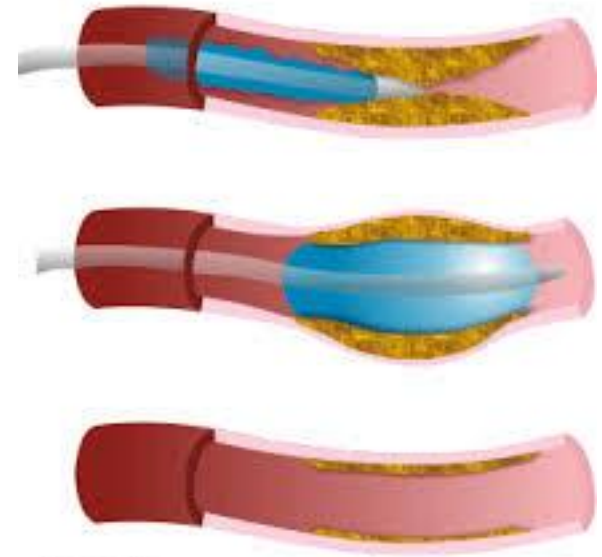
thrombolyse

# Contexte de l'étude

→ Infarctus du myocarde



90 min



percutaneous coronary intervention

# Contexte de l'étude

## → Épidémiologie AVC et infarctus du myocarde au Sénégal

### › Taux d'Incidence AVC:

100 pour 100 000 habitants (~13 500 cas/an)

*the incidence of stroke in Europe is between 63 and 159 for 100000 women, and between 101 and 239 for 100000 men*

### › Taux d'Incidence de infarctus du myocarde :

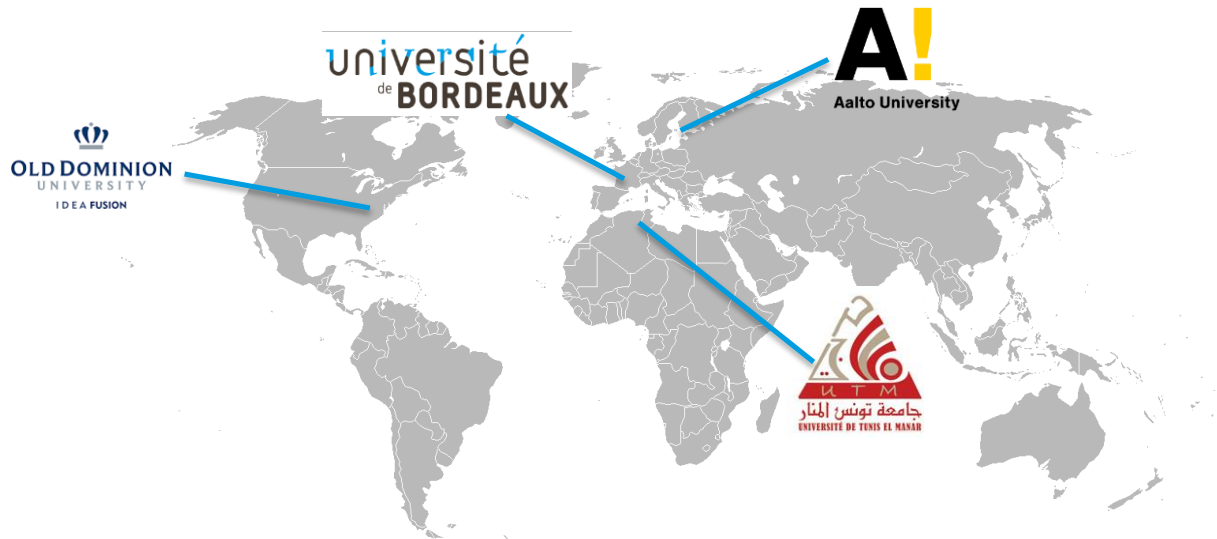
150 pour 100 000 habitants (~24 300 cas/an)

*180 per 100000 inhabitants in France*

## → Nombre de cas directement proportionnel à la densité de population

## → Combien de patients ne peuvent pas atteindre l'hôpital à temps?

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



- <sup>a</sup> *Faculté des Sciences de Tunis, University of Tunis, Tunisia*
- <sup>b</sup> *ERIAS INSERM U897, ISPED, University of Bordeaux, F-33000, France*
- <sup>c</sup> *Virginia Modeling Analysis and Simulation, Old Dominion University, VA, USA*
- <sup>d</sup> *Department of Communications and Networks, Aalto University School of Electrical Engineering, Espoo, Finland*

# Data used

- Anonymised Dataset provided by Orange and Senegal context data
- General Population and Housing Sensus data (2013)
- Guidelines of treatments from French HAS and incidence rates from scientific literature
- Hospital data from *SenDoctor* and *Senegal Medical Directory*
- Maps of Senegal (OpenStreetMap)



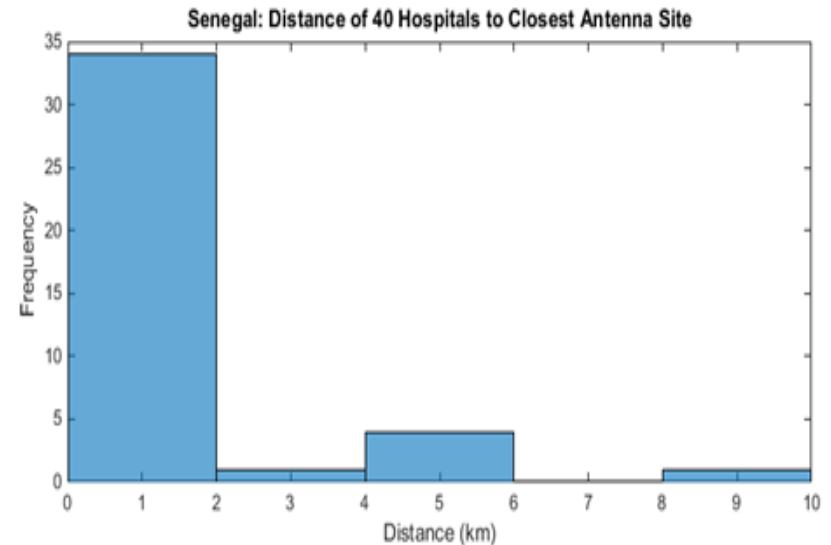
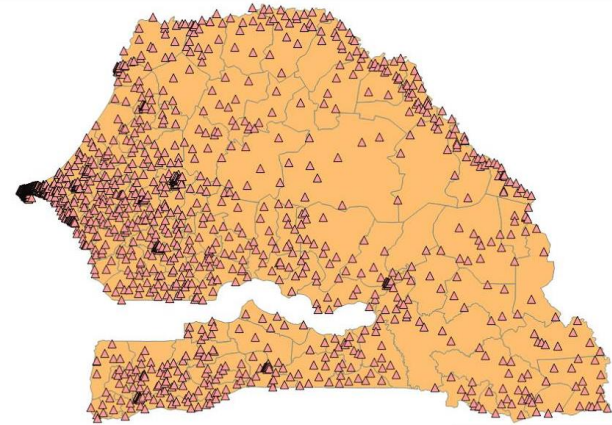
# Approach

- L'évaluation de la distance et du temps pour atteindre l'hôpital le plus proche
- Calcul des taux densité de population + d'incidence dans le secteur de l'antenne téléphonique
- Identification des zones à risque (trop éloignées d'un centre) et estimation des cas non évités

# Approach (cont.)

## → Distance Evaluation

- › **Objective:** to evaluate distance from each location in Senegal to nearest hospital
- › Geographical location of the hospitals approximated by nearest antenna site
  - Under this approximation 85% of the 40 hospitals considered where within 2 km of their real geographical locations
  - The impact of this error is minimal when evaluating the travel time to the hospitals.

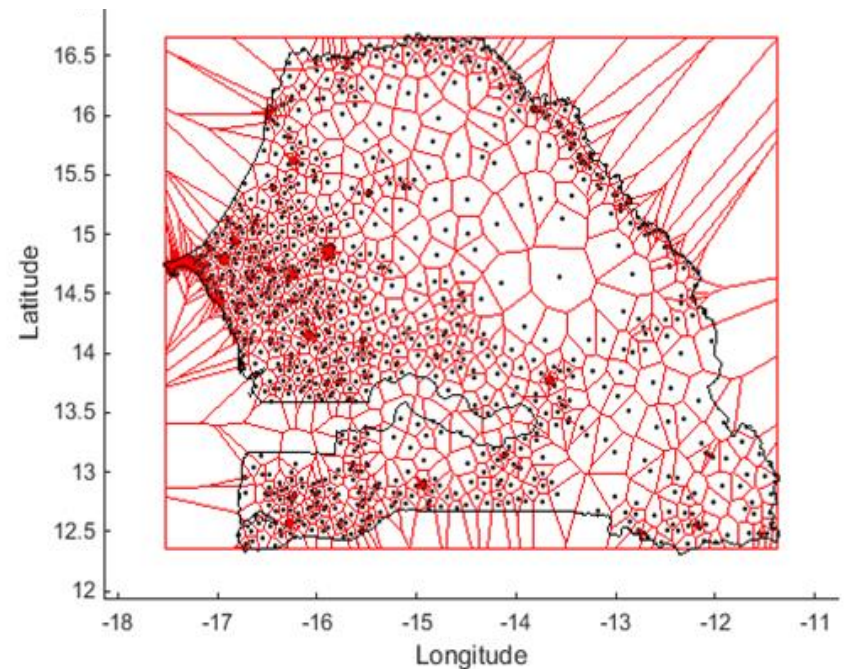


# Approach (cont.)

## → Distance Evaluation

- › Different geographical locations are then app segmented according to antenna coverage areas
- › Coverage area evaluated using Voronoi tessellations
- ›
- › Antenna site location assumed to be default location for whole the coverage area

Voronoi cell layout for Senegal based on provided 1666 site locations



# Approach (cont.)

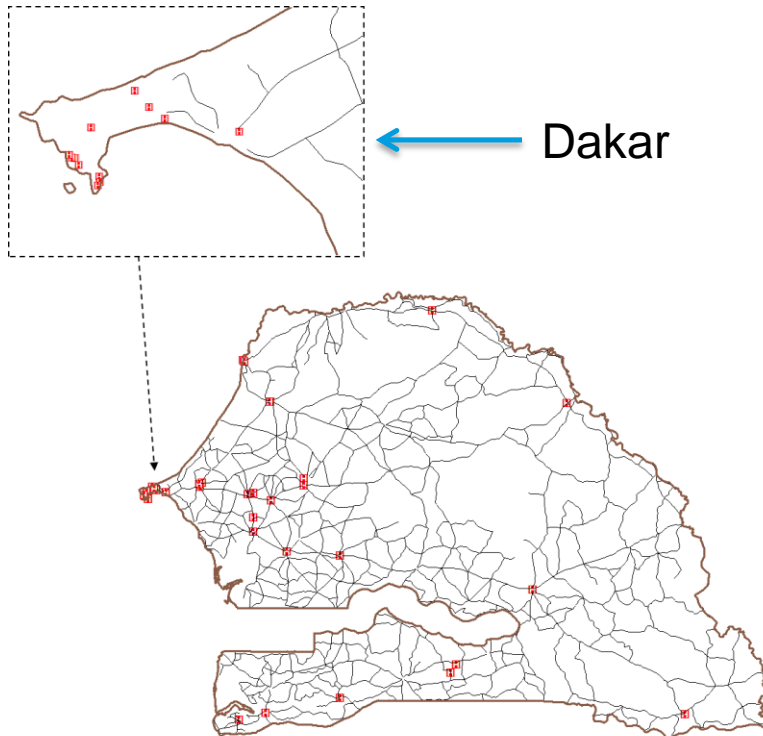
## → Distance Evaluation

- › Distance is then evaluated as straightline (Euclidean) distance between a coverage area (area's antenna location) to nearest hospital (also represented as an antenna location)
  - Correction factor known as **detour index** is applied on distance to account for actual route of road network (road density), road quality, road congestion control, etc.
  - Detour index of 1.2-1.6 usually assumed in countries with well developed road infrastructure
  - For Senegal a more conservative detour index of 2.0 is used

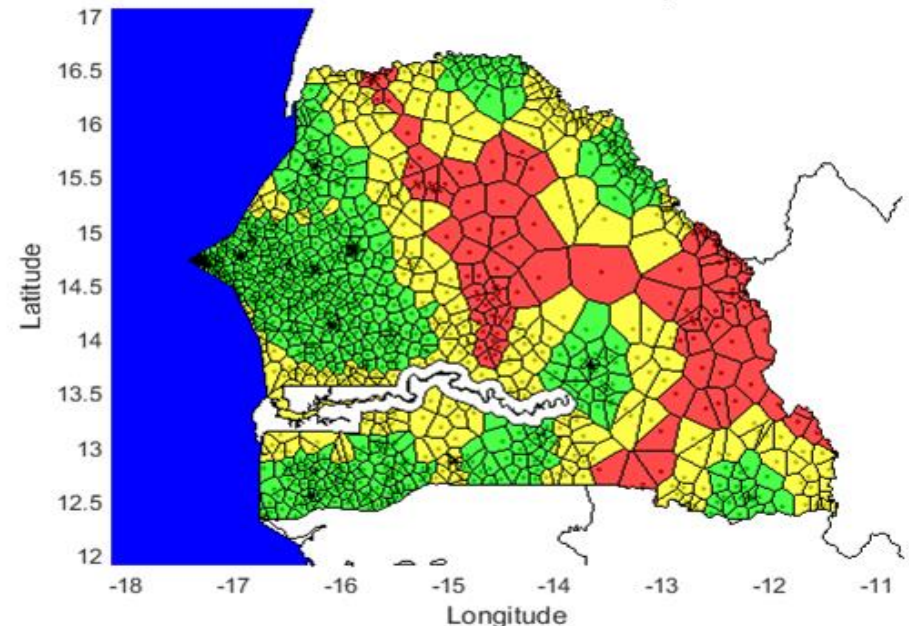
# Approach (cont.)

## → Estimating travel times to reach the nearest hospital

Location of the 40 hospitals considered in the study



Travel Times to nearest Hospital

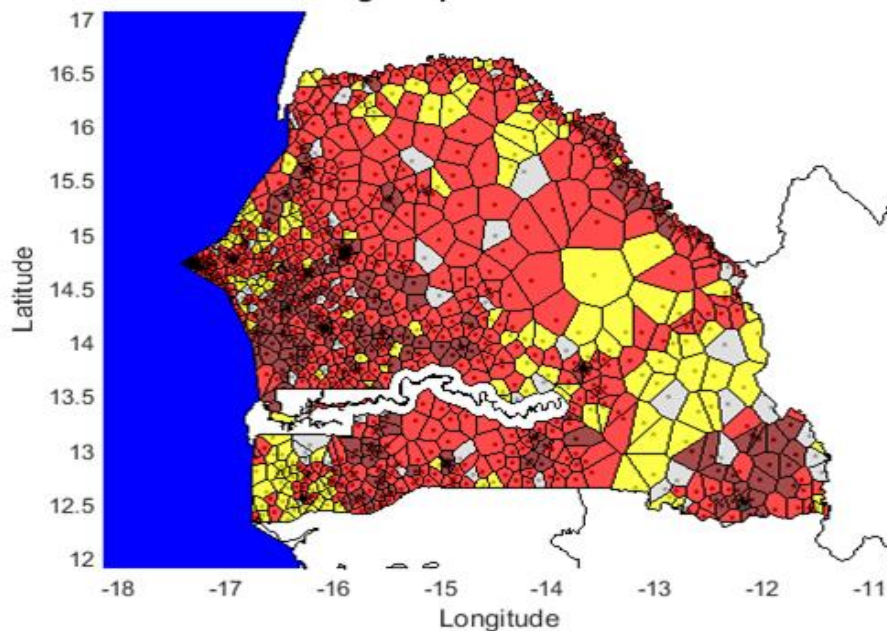


Green: less than 90 minutes,  
Yellow: 90 to 180 minutes,  
Red: over 180 minutes

# Approach (cont.)

→ Computing the population density + incidence rates at antenna (site location) level

Distribution of Senegal Population according to the mobile antenna



grey  $N < 100$ ,  
yellow for  $101 < N < 1000$ ,  
red for  $1001 < N < 10000$ ,  
brown for  $10001 < N < 100000$   
black for  $N > 100000$

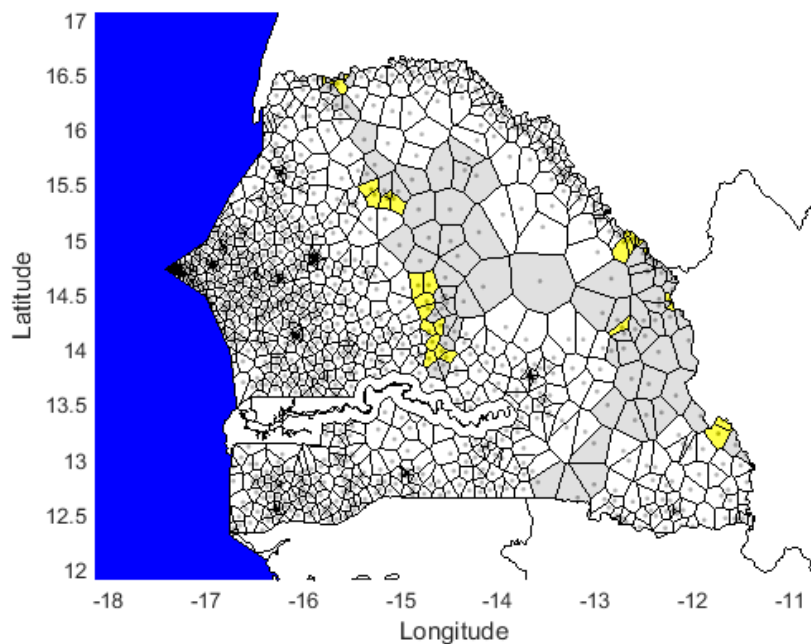
## Correction Factors

$$\alpha = U_S * 1.2 * 1 / O_{ms}$$
$$\beta = \frac{U_R}{O_R}$$

# Approach (cont.)

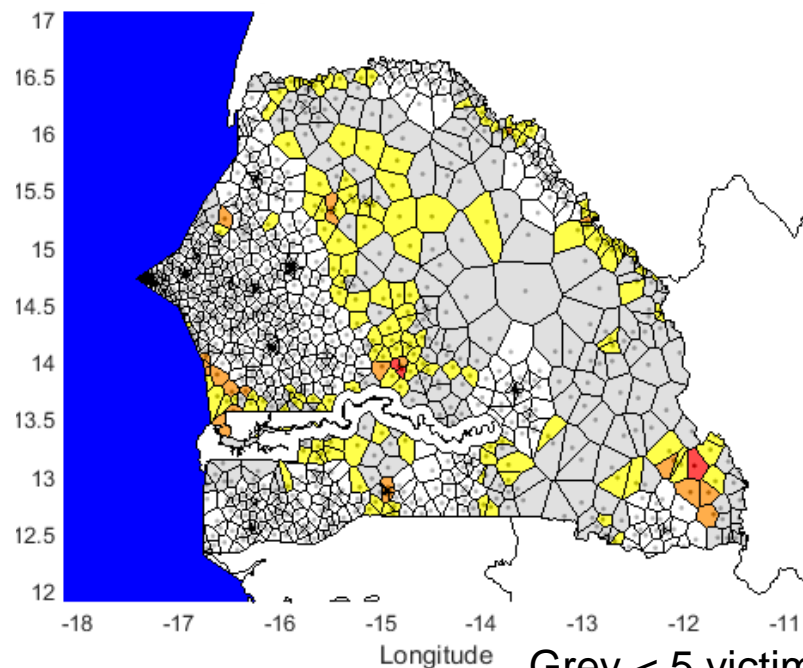
→ Highlighting white zones: areas where people are at high risk in case of Stroke or Myocardial Infarction

Estimated incident cases of strokes from different areas (432)



Grey < 5 victims,  
Yellow 5 to 25

Estimated incident cases of Myocardial Infarction from different areas (4241)



Grey < 5 victims,  
Yellow 5 to 25,  
Orange 26 to 50,  
Red 51 to 100

# Limites de l'étude

- Biais lié à l'extrapolation de la population dans une zone de couverture de l'antenne donnée
  - › Comme si chaque personne avait un téléphone cellulaire
  - › Comme si Orange a été le seul opérateur de téléphonie mobile au Sénégal
  
- Un filtrage des données est effectuée par Orange
  
- L'étude est basée sur un taux d'incidence estimé des pathologies considérées (pas de chiffres officiels récents pour le Sénégal)



# Conclusion

- Our findings suggest that the analysis and cross linking of big anonymized mobile dataset helps making prediction
- The identification of areas at high risk in case of stroke of myocardial infarction in Senegal could help Public Health decision makers to take the required actions on the earlier
- Future Work
  - › Risk estimation based on more fine-grained population density estimation
  - › Introducing parameters on the hospital capabilities and additional contextual conditions (Linked Open Data exploitation)
  - › Taking into account other emergency cases, relying on Semantic Web technologies and Ontology Modelling