TYPOLOGY OF HEALTH DISTRICTS BASED ON DATA FROM TEN INFECTIOUS DISEASES UNDER SURVEILLANCE IN SENEGAL

Abdourahmane Ndao and Cheikh Tidiane Seck[*]

Département of Mathematics, Alioune Diop University, Bambey, Senegal

[*]Corresponding author: cheikhtidiane.seck@uadb.edu.sn

ABSTRACT: *The aim of this work is to construct a typology of health districts in Senegal based on the distribution frequency of ten infectious diseases under surveillance. Our methodology utilizes HCPC (Hierarchical Classification on Principal Components) algorithm which combines two data analysis techniques, namely Principal Component Analysis (PCA) and Hierarchical Ascending Classification (HAC). The data come from the Prevention Department of the Ministry of Health and Social Action and cover the period from January 2018 to November 2022. The results show that health districts in Senegal can be divided into three clusters according to the number of confirmed cases recorded for each of the ten considered infectious diseases. Moreover, the parangons' principle allows us to select from the obtained clusters a representative stratified sample of health districts in view to identifying risk factors associated with these ten pathologies.*

KEYWORDS: Typology; Health districts; Number of confirmed cases; Infectious diseases; Health surveillance data.

INTRODUCTION

In the context of re-emergence of certain infectious diseases, epidemiological surveillance is crucial for safeguarding populations' health. It is a continuous process of collecting, compiling, analyzing and disseminating data on pathologies. It is essential for planning, implementing and evaluating health practices. It is thus a precious tool for defining health policy and programs. In Senegal, health authorities have recently noted a resurgence of some infectious diseases that appeared to be eradicated in the past. This is why the government has developed a national technical guide for the integrated disease surveillance and response (SIMR), which is a reference document for all the epidemiological surveillance activities at any level of the health pyramid. This document provides guidelines for re-enforcing early detection and response capacity to face epidemics and other health emergencies.

The surveillance data collection and analysis system has two main components: the indicator-based surveillance (SBI) and the event-based surveillance (SBE). SBI focuses on structured data collected from identified formal sources such as hospitals, health centers, health posts, etc. whereas SBE collects unstructured data, which may be signals or reports on a potential public health risk, transmitted either formally (via health facilities) or informally (via the media).

However, it is worth noting that the collected surveillance data are often incomplete and may have errors. Hence, it is necessary to correct and complete them in order to carry out reliable statistical analyses for decision-making purposes.

In public health, typologies are often used to understand and better manage the health of populations. For example, [1] used a cluster sampling method to introduce a new typology of long-term care (LTC) systems in OECD member countries, based on the latest data and a single set of institutional indicators. [8] used a series of cluster analyses to develop an expanded typology of OECD health systems. The authors found that at least five distinct types of health systems can be identified and provide quantitative information on the coherence of cluster membership for each country. [2] used classification technique to examine evidence from recent studies on the complexity of diabetes by proposing 4 sub-groups (sub-types) of diabetes: - severe insulin-deficient diabetes; - insulin-diabetes severe resistance; - diabetes related to mild obesity; and - diabetes related to mild age. [6] conducted a cluster analysis to study and construct a typology of general practitioner's perceptions of vaccines that caused controversy in France in 2014. The analysis identified four groups of physicians based on their sensitivity to vaccine controversies, and factors associated with certain interest groups were identified using logistic regression models.

In this paper, we analyze data coming from the integrated surveillance of ten infectious diseases in Senegal: *COVID-19, Measles, Poliomyelitis (PFA), Dengue, Meningitis, Rift Valley Fever (Rift), Lassa and Crimean-Congo Hemorrhagic Fever (CCHF), Chikungunya (CHIK), West Nile Virus Fever (WN) and Yellow Fever (YF)*. The data concern the number of confirmed cases recorded for each of the 10 diseases and each of the 79 health districts in Senegal. Our aim is to construct a typology of these 79 health districts using the distribution frequency (i.e., the number of confirmed cases) of these 10 infectious diseases. The choice of these 10 diseases can be justified by the importance given by Senegalese health authorities to their control and monitoring. This typology will make it possible to locate where these different infectious diseases are spreading. Moreover, it can be used as a first stratification step to select a

representative sample of health districts in which we should investigate the direct risk factors associated with these ten infectious diseases.

The remainder of the paper is organized as follows: In Section 2, we present the data and review the main technique used to produce the typology; Section 3 presents the results; Section 4 provides a discussion, while Section 5 concludes the paper.

METHODOLOGY

Data

The dataset comes from the integrated surveillance of 10 infectious diseases that are prevalent throughout the country. It covers the period from January 2018 to November 2022 and contains 113,847 suspected cases for all of the 10 diseases. The suspected cases are collected in each health district (DS) considered as an epidemiological unit. From this dataset, we filtered the number of confirmed cases corresponding simultaneously to each health district and to each infectious disease. This leads to an *individuals × variables* table providing the cross number of confirmed cases for each of the 10 targeted diseases and each of the 79 health districts. Table 2.1 in the appendix gives an insight into the data. Our aim is to build a typology of the health districts, based on the number of confirmed cases of these diseases.

Method

We use the HCPC (Hierarchical Classification on Principal Components) algorithm (see [3], [4]), which combines two data analysis techniques: Principal Component Analysis (PCA) and Hierarchical Ascending Classification (HAC). The algorithm performs the following steps below, and allows to obtain a stratified sample from the initial set of items to be classified:

1. Perform a PCA on the extracted dataset;

2. Apply hierarchical ascending classification to the PCA result;

3. Draw the classification tree (or dendrogram) (see Figure 3.1);

4. Plot the resulting clusters with their constitutive elements (see Figure 3.2);

5. Create a class variable on the initial dataset;

6. Characterize the clusters by comparing the average in the category (i.e., the cluster) with the overall average over all clusters, or by examining the p-values associated with the different diseases (see Table 3.2);

7. Select the most representative elements in each cluster, based on the *parangons'* principle (see Table 3.3).

RESULTS

Description of the Target Variable (Number of Confirmed Cases)

Our variable of interest is the number of confirmed cases of each of the 10 infectious diseases considered. Recall that a confirmed case is established by laboratory diagnostic tests performed within the districts. An epidemic threshold is fixed for each infectious disease and enables one to say whether there is an epidemic (see Table 3.2). The aim is to construct a typology of the health districts, based on the number of confirmed cases recorded for each of the 10 infectious diseases. Table 3.1 shows the number of confirmed (positive test result) and non-confirmed (negative test result) cases among all the 113,847 suspected cases recorded overall the ten diseases.

Table 3.1 : Frequency of positive and negative cases

	Positive	Negative	Total
Count	68 698	45 149	113 847
Proportion (%)	60	40	100

Table 3.2 shows the distribution of suspected cases (CasSusp) and confirmed cases (CasConf) according to the ten infectious diseases considered, and indicates that COVID-19, Measles, Dengue and Meningitis are the most frequently distributed diseases during the covered period.

Table 3.2 : Distribution of suspected and confirmed cases and epidemic threshold

Disease	CasSusp	CasConf	Epidemic threshold	Alert
CCHF	5 729	11	1	Only one confirmed case to declare an epidemic
CHIK	5 729	4	1	Only one confirmed case to declare an epidemic
Covid-19	68 817	66 907	1	Only one confirmed case of Covid-19 = epidemic
Dengue	5 729	291	1	Only one confirmed case of Dengue to declare an epidemic
Meningitis	5 570	260	1	A single confirmed case of Meningitis = an epidemic
PFA	1 067	18	1	A single confirmed case of poliomyelitis to declare an epidemic
Rift	5 729	13	1	One confirmed case = epidemic
Measles	4 019	1 115	3 CasSusp/ CasConf	3 (or more) suspected cases observed in a district or health
WN	5 729	42	1	a month are laboratory confirmed (lgM+)
YF	5 729	37	1	One confirmed case of WN = an epidemic
Sum	113 847	68 698		

Figure 3.1 illustrates the distribution (in percentage) of all confirmed cases according to the ten considered diseases. This corresponds to the COVID-19 pandemic period, which accounted for almost all the confirmed cases recorded, that is, 97.39% of the confirmed cases.
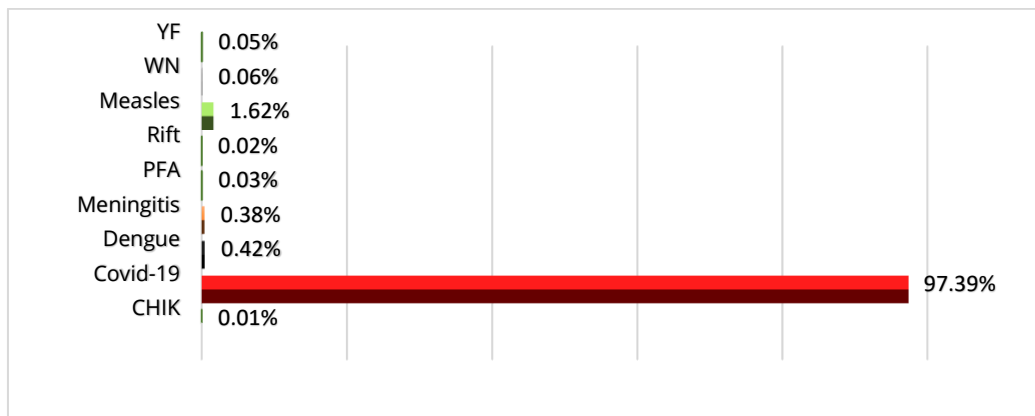

Figure 3.1: Distribution frequency of the 10 infectious diseases

Interrelationships between the Ten Infectious Diseases

A principal component analysis (PCA) is performed on Table 2.1 and reveals that there are some correlations between the ten infectious diseases. On one hand, there is a positive correlation between the numbers of confirmed cases of WN, Meningitis, YF, CHIK and COVID-19. On the other hand, the numbers of confirmed cases of Rift Valley Fever and CCHF are positively correlated with each other, but are both negatively correlated with the number of confirmed cases of Measles (see Figure 3.2).
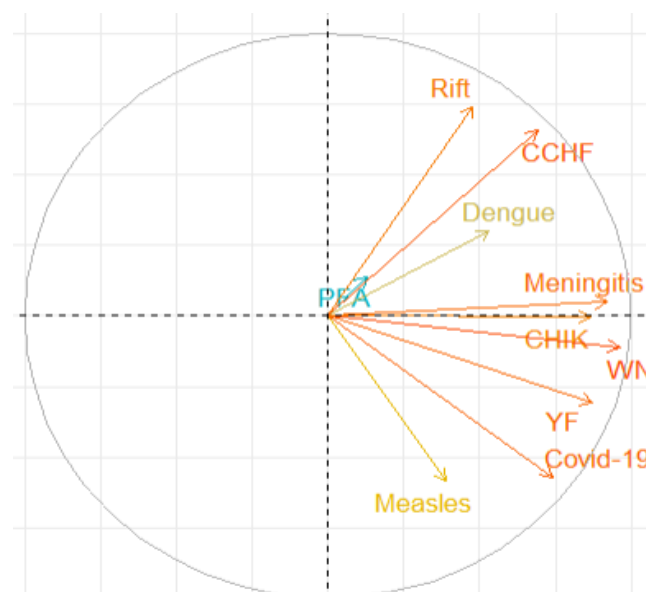

Figure 3.2: Correlations between the ten infectious diseases

Construction of the Typology

Figure 3.3 shows the health districts' classification tree resulting from the application of the HCPC algorithm, which suggests an optimal cut-off level materialized by a horizontal line (here at height of around 1.5).
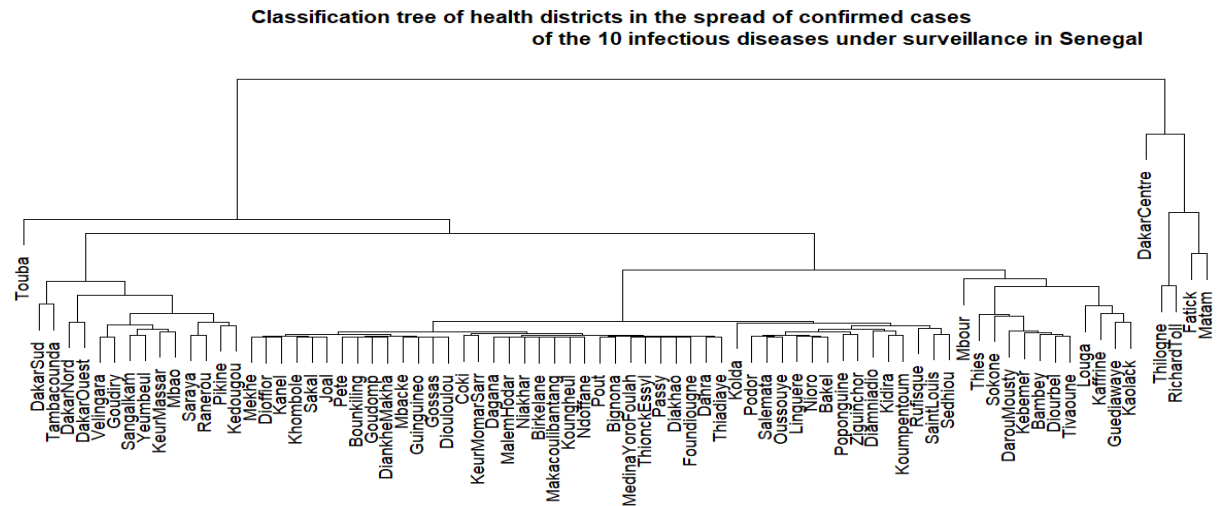

Figure 3.3: Classification tree of Senegal's health districts

This classification tree shows that the health districts in Senegal may be grouped into three clusters. Figure 3.4 illustrates a typology map of the health districts, where each cluster is represented with its own members in a different color. The first cluster, comprising the majority of the districts, is shown in blue. The second cluster represented in yellow comprises the districts of Richard Toll, Thilogne, Fatick and Matam while the third cluster shown in grey only comprises the district of Dakar-Centre.
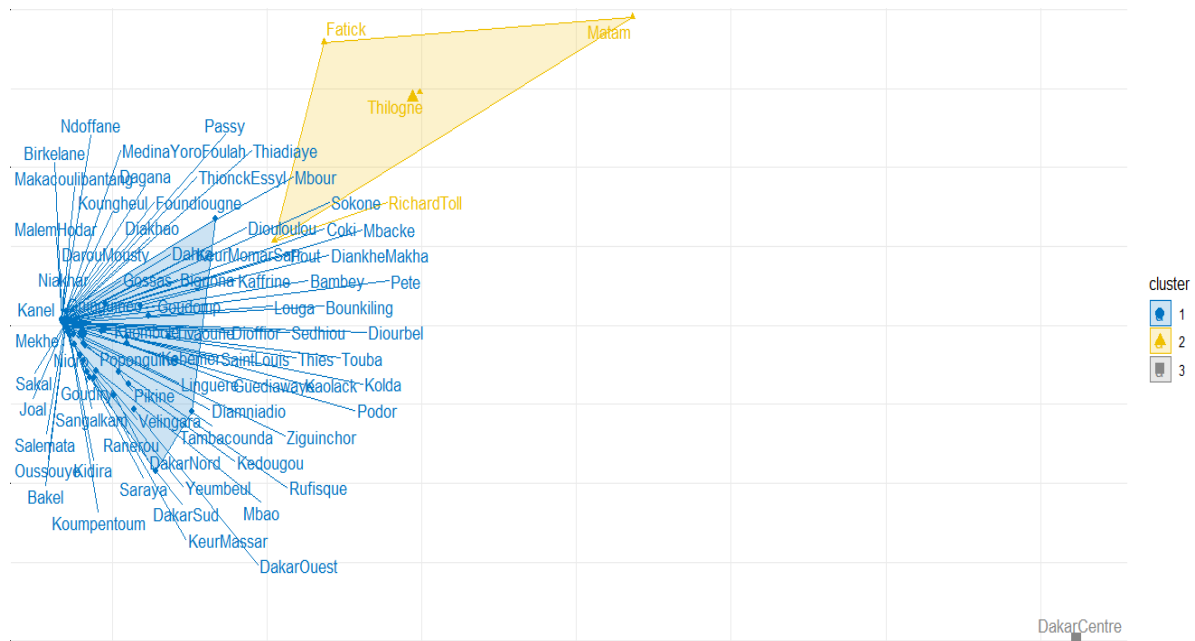

Figure 3.4: Typology of health districts

Description of the Clusters

Table 3.3 presents the results of the description of each of the three clusters by the ten infectious diseases considered. It shows that Yellow Fever (YF), COVID-19, Meningitis, West Nile Fever (WN) and Measles are significantly present in Cluster 3, and thus characterize this cluster, as the average number of confirmed cases for each of these diseases in this cluster is greater than the overall average number of confirmed cases for all clusters.

One can note that the diseases – Dengue, Meningitis, Rift Valley fever (Rift), Lassa hemorrhagic fever and Crimean-Congo hemorrhagic fever (CCHF) – are all present in the health districts of cluster 2. However, this cluster is more characterized by Rift Valley Fever and CCHF diseases, which are only present in the districts of this cluster.

Finally, Cluster 1, which includes the majority of the health districts, is not significantly characterized by any of the ten infectious diseases considered, although there is a low prevalence of some pathologies such as: COVID-19, Dengue and Meningitis.

Table 3.3 : Results for clusters' characterization

Cluster	Infectious diseases	Mean in category	Overall mean	p.value
1	COVID-19	578.36	846.76	2.583830e-04
	Dengue	2.16	3.68	8.290052e-06
	YF	0.15	0.47	1.726225e-07
	CHIK	0.00	0.05	7.121711e-10
	WN	0.12	0.53	2.033987e-11
	Rift	0.01	0.16	1.550241e-12
	Meningitis	1.67	3.28	9.927304e-13
	CCHF	0.04	0.14	3.330041e-13
2	Rift	4.00	0.16	1.912934e-16
	CCHF	2.33	0.14	1.520681e-13
	Meningitis	25.66	3.28	5.492230e-06
	Dengue	32.67	3.68	9.991097e-05
	CHIK	0.67	0.05	5.802867e-04
	WN	5.00	0.53	8.415831e-04
3	YF	20	0.47	1.057915e-16
	COVID-19	22 891	846.76	4.256673e-15
	WN	18	0.53	1.018230e-13
	Meningitis	57	3.28	5.120433e-10
	CHIK	2	0.05	5.475368e-10
	Measles	90	14.11	4.116857e-04
	Dengue	31	3.68	3.663010e-02

Selection of the Most Representative Health Districts

To select the most representative health districts in view to identifying risk factors associated with these ten infectious diseases, we use the *parangons'* principle (the closest element to the center of each cluster). By default, the HCPC algorithm selects a number of *parangons* equal

to 5 *(nb.par = 5)*. So, we will determine the first 5 *parangons* of Cluster 1, the size of which is greater than 5. Then, since the other two clusters have sizes less than 5, we will consider all their members as *parangons.*

Table 3.4 shows the selected health districts with their distances from the cluster centers. This gives a stratified sampling procedure in view to identifying risk factors associated with these infectious diseases.

Table 3.4 : Selection of the most representative health districts (DS)

Cluster	Most representative DS	Distance	Number of DS selected
1	Koumpentoum	0.43	$n_1 = 5$
	Kidira	0.45	
	Linguere	0.46	
	Diamniadio	0.48	
	Nioro	0.51	
2	Fatick	3.65	$n_2 = 4$
	Thilogne	4.89	
	Matam	5.36	
	Richard Toll	5.55	
3	DakarCentre	0	$n_3 = 1$
Total			10

Table 3.5 shows the link between the most representative health districts in each cluster and the infectious diseases that characterize that cluster. Thus, in Cluster 2, which includes the districts of Fatick, Thilogne, Richard-Toll and Matam, the survey to identifying risk factors will focus on Dengue, Rift Valley Fever (Rift) and Lassa and Crimean-Congo Hemorrhagic Fever (CCHF).

Table 3.5 : Link between the most representative health districts and infectious diseases

Infectious diseases	Cluster	Representative health districts
COVID-19 CHIK YF Rift WN Meningitis Dengue CCHF	1	Koumpentoum Kidira Linguere Diamniadio Nioro
Dengue Rift CCHF	2	Fatick Thilogne Richard Toll Matam

COVID-19 Measles YF Meningitis WN CHIK	3	DakarCentre

DISCUSSION

The aim of this work was to construct homogeneous classes of health districts based on the distribution frequency of ten infectious diseases. It is an analytical and prospective study covering a 4-year period from January 2018 to November 2022. We obtained a typology of the health districts in Senegal by dealing a dataset extracted from the global surveillance database provided by the Prevention Department of the Ministry of Health and Social Action.

The study reveals well-defined areas of spread for each of the 10 infectious diseases. Indeed, we find that COVID-19, Yellow Fever, Meningitis, West Nile Fever (WN) and Measles significantly characterize Cluster 3, which only comprises Dakar-Centre district. Hence, Dakar-Centre can be considered as a hotbed for the spread of these five infectious diseases.

As well, Cluster 2, including the health districts of Fatick, Thilogne, Richard Toll and Matam (see Table 3.3), is significantly characterized by four infectious diseases: Dengue, Rift, CCHF and Meningitis. As a consequence, the hotbed of these four diseases can be confined to these health districts.

CONCLUSION

The purpose of this study was to construct a typology of health districts in Senegal in order to better locate the sources of spread of ten infectious diseases across the country. The HCPC classification method was utilized and led to three clusters of health districts; each one has been characterized by well-defined infectious diseases. Moreover, the study enabled us to select a representative stratified sample of 10 health districts to be surveyed in view to identifying direct risk factors associated with these pathologies. The selected health districts are: Koumpentoum, Kidira, Linguère, Diamniadio, Nioro, Fatick, Thilogne, Richard-Toll, Matam and Dakar-Centre.

The major interest of this study lies in the fact that it constitutes a preliminary step to an undergoing work which consists of determining direct risk factors associated with these ten infectious diseases, in the sense that it narrows the field of analysis and facilitates the preparation of questionnaires for collecting additional information pertaining to these infectious diseases.

REFERENCES

[1]. Ariaans M, Linden P, Wendt C. Worlds of long-term care: A typology of OECD countries. Health Policy. 2021 May;125(5):609-617. doi: 10.1016/j.healthpol.2021.02.009. Epub 2021 Feb 25. PMID: 33715875. DOI : 10.1016/j.healthpol.2021.02.009

[2]. Herder C, Roden M. A novel diabetes typology: towards precision diabetology from pathogenesis to treatment. Diabetologia. 2022 Nov;65(11):1770-1781. doi: 10.1007/s00125-021-05625-x. Epub 2022 Jan 4. PMID: 34981134; PMCID: PMC9522691.

[3]. Husson F., Josse J., and Pagès J. (2010). "Principal Component Methods - Hierarchical Clustering - Partitional Clustering: Why Would We Need to Choose for Visualizing Data?" *Unpublished*

[4]. Husson F., Sebastien L., and Pagès J. (2017). *Exploratory Multivariate Analysis by Example Using R.* 2nd ed. Boca Raton, Florida: Chapman, Hall/CRC.

[5]. Legba C. S., Yabi, I., Azonhe T., Osseni A. A. (2017). Analyse Des Déterminants Du Choléra Dans La Ville de Djougou Au Bénin. *European Scientific Journal*, June 2017 edition, Vol.13, No.18. Doi: 10.19044/esj.2017.v13n18p171.

[6]. Le Marechal M, et al., General practitioners' perceptions of vaccination controversies: a French nationwide cross-sectional study, *Clinical Microbiology and Infection (2017)*, https://doi.org/10.1016/j.cmi.2017.10.021.

[7]. MSAS/DP: Guide technique nationale pour la surveillance intégrée de la maladie et la riposte. 3ᵉ Edition, Voulme 1- Section Introduction, Décembre 2019.

[8]. Munn, Z., Stern, C., Aromataris, E. *et al.* What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Med Res Methodol* **18**, 5 (2018). https://doi.org/10.1186/s12874-017-0468-4

[9]. Reibling N, Ariaans M, Wendt C. Worlds of Healthcare: A Healthcare System Typology of OECD Countries. Health Policy. 2019 Jul;123(7):611-620. doi: 10.1016/j.healthpol.2019.05.001. Epub 2019 May 11. PMID: 31133444.

[10]. Serge M., Agnès W-S. (2012), Les déterminants des maladies infectieuses humaines en Europe : influences de la biodiversité et de la variabilité climatique. BEH 12-13 / 20 mars 2012.

APPENDIX

Table 2.1: Extract from the dataset

	Measles	PFA	Meningitis	Dengue	Covid-19	Rift	WN	CHIK	YF	CCHF
DakarCentre	90	0	57	31	22891	0	18	2	20	1
DakarNord	38	0	0	0	4640	0	0	0	0	0
DakarOuest	25	0	0	0	8233	0	0	0	0	0
DakarSud	102	0	0	0	5888	0	0	0	0	0
Diamniadio	13	0	0	0	707	0	0	0	1	0
Guediawaye	19	0	10	10	2075	0	0	0	0	0
KeurMassar	46	0	0	0	1335	0	0	0	0	0
Mbao	36	0	1	0	1822	0	0	0	0	0