Application of functional data analysis to classify communities according to the pattern of malaria incidence, to guide targeted control strategies.

Sokhna DIENG (sokhna.dieng@etu.univ-amu.fr)
Aix Marseille Univ, IRD, INSERM, SESSTIM, Marseille, France https://orcid.org/0000-0003-0603-1986

Pierre Michel
Aix Marseille Univ, CNRS, EHESS, Centrale Marseille, AMSE, Marseille, France

Abdoulaye Guindo
Aix Marseille Univ, IRD, INSERM, SESSTIM, Marseille, France

Kankoe Sallah
Aix Marseille Univ, IRD, INSERM, SESSTIM, Marseille, France

El-hadj Ba
UMR VITROME, Campus International IRD-UCAD de l'IRD, Dakar, Senegal

Badara Cissé
Institut de Recherche en Santé, de Surveillance Epidemiologique et de formation (IRESSEF), Diamniadio, Senegal

Patrizia Carrieri
Aix Marseille Univ, IRD, INSERM, SESSTIM, Marseille, France

Cheikh Sokhna
UMR Vitrome, Campus International IRD-UCAD de l'IRD, Dakar, Senegal

Paul Milligan
London School of Hygiene and Tropical medicine, London, United Kingdom

Jean Gaudart
Aix Marseille Univ, APHM, INSERM, IRD, SESSTIM, Hop Timone, BioSTIC, Biostatistic & ICT, Marseille, France

---

Research article

Keywords: Functional data analysis, Time series clustering, Malaria patterns, Malaria dynamic, Hierarchical ascending clustering,

Posted Date: February 26th, 2020

DOI: https://doi.org/10.21203/rs.2.24479/v1

Version of Record: A version of this preprint was published on June 11th, 2020. See the published version at https://doi.org/10.3390/ijerph17114168.

Application of functional data analysis to classify communities according to the pattern of malaria incidence, to guide targeted control strategies.

Sokhna Dieng*[1], Pierre Michel[2], Abdoulaye Guindo[1, 3], Kankoe Sallah[1, 4], El-hadj Ba[5], Badara Cissé[6], Patrizia Carrieri[1], Cheikh Sokhna[5], Paul Milligan[7], Jean Gaudart[8]

[1] Aix Marseille Univ, IRD, INSERM, SESSTIM, Marseille, France.

[2] Aix Marseille Univ, CNRS, EHESS, Centrale Marseille, AMSE, Marseille, France

[3] Malaria Research and Training Center - Ogobara K Doumbo, FMOS-FAPH, Mali-NIAID-ICER, Université des Sciences, des Techniques et des Technologies de Bamako, Bamako, Mali

[4] AP-HP, Hôpital Bichat, Unité de Recherche Clinique PNVS

[5] UMR VITROME, Campus International IRD-UCAD de l'IRD, Dakar, Senegal

[6] Institut de Recherche en Santé, de Surveillance Épidémiologique et de Formation (IRESSEF), Diamniadio, Sénégal

[7] London school of hygiene and tropical medicine, London, United Kingdom

[8] Aix Marseille Univ, APHM, INSERM, IRD, SESSTIM, Hop Timone, BioSTIC, Biostatistic & ICT, Marseille, France

* Corresponding author: sokhna.dieng@etu.univ-amu.fr

Abstract

Background

Effective targeting of malaria control in low transmission areas requires identification of transmission foci or hotspots. We investigated the use of functional data analysis to identify and describe spatio-temporal pattern of malaria incidence in an area with seasonal transmission in west-central Senegal.

Method

Malaria surveillance was maintained over 5 years from 2008 to 2012 at health facilities serving a population of 500,000 in 575 villages in two health districts in Senegal. Smooth functions were fitted from the time series of malaria incidence for each village, using cubic B-spline basis functions. The resulting smooth functions for each village were classified using hierarchical clustering (Ward's method), using several different dissimilarity measures. The optimal number of clusters was then determined based on four cluster validity indices, to determine the main types of distinct temporal pattern of malaria incidence. Epidemiological indicators characterizing the resulting malaria incidence pattern in terms of the timing of seasonal outbreaks, were calculated based on the slope (velocity) and rate of change of the slope (acceleration) of the incidence over time.

Results

Three distinct patterns of malaria incidence were identified. A pattern characterized by high incidence, in 12/575 (2%) villages, with average incidence of 114 cases/1000 person-years over the 5 year study period; a pattern with intermediate incidence in 97 villages (17%), with average incidence of 13 cases/1000 person-years; and a pattern with low incidence in 466 (81%) villages, with average incidence 2.6 cases/1000 person-years. Epidemiological indicators

characterizing the fluctuations in malaria incidence showed that seasonal outbreaks started later, and ended earlier, in the low incidence pattern.

Conclusion

Functional data analysis can be used to classify communities based on time series of malaria incidence, and to identify high incidence communities. Indicators can be derived from the fitted functions which characterize the timing of outbreaks. These tools may help to better target control measures.

Trial registration

The data used in this work come from the clinical trial NCT 00712374. The trial is registered at www.clinicaltrials.gov, number NCT 00712374.

Keywords: Functional data analysis; Time series clustering; Malaria patterns; Malaria dynamic; Hierarchical ascending clustering;

Background

In areas of low malaria transmission, because of the spatial heterogeneity of malaria incidence, WHO recommends the development of targeted control strategies adapted to the local epidemiological context [1] . Effective targeting requires identification of transmission foci or hotspots based on epidemiological and environmental data. In this paper we introduce an approach using functional data analysis to classify villages on the basis of the pattern of malaria incidence over a 5-year period, to inform the development of targeted control strategies in west-central Senegal.

The principle of functional data analysis is that the observed data can be described by smooth functions which are estimated in order to understand the underlying patterns in the data. These functions can be obtained for each of a large number of units (communities), clustering algorithms can then be used to identify broad types of temporal pattern. For time series of disease incidence, the slope of the function, and the rate of change in the slope (the velocity and acceleration), are useful indicators which can be calculated to characterize key features of the data. A considerable amount of research have been dedicated to the development of statistical methods and tools for analysis of functional data [2–6]. The work by Ramsay *et al.* has made these approaches popular, and R and Matlab programs have made the methods available to a wider group of researchers [5]. Applications in public health and biomedical sciences are reviewed by Ullah and Finch (2013) [7]. The aim of this study was to investigate the use of these methods to understand spatio-temporal variation of malaria incidence, to inform targeting malaria control measures in a low transmission area of Senegal. Existing approaches used to target malaria risk areas are based on incidence rate, and prevalence [8–14] in discrete time periods [15–19]. In contrast, the functional data approach the whole time series to be characterized, potentially giving a better understanding of how intervention should be targeted.

Methods

Study area and dataset

The data used for this study from 2008 to 2012 were collected during a field trial of Seasonal Malaria Chemoprevention in children in the west-central Senegal [20, 21]. Data for 575 villages in the rural health districts of Bambey and Fatick, have been used in this analysis. This area is a part of the two national health district where national malaria control estimated the incidence under 5 cases /1000 person-years in 2018 [22].

Malaria surveillance was maintained in 38 health facilities serving a population of about 500,000 living in 575 villages (single villages or groups of adjacent hamlets). Malaria cases were patients at health facilities with fever or history of fever, in the absence of evident alternative causes of the fever, who had a positive rapid diagnostic test (RDT). Date of diagnosis, village of residence, and other details were obtained for each case from facility registers. The surveillance system is described by Cisse et al (in the main paper and in supplement S3) [20]. The population was enumerated through a census in 2008 and updated through visits to each household at approximately 10 months intervals from 2008 to 2012. Coordinates of the centre of each village were obtained by GPS.

For this analysis only aggregated data by villages were used.

Statistical analysis

-Determining the smooth function (functional data) for each time series

The time series of the observed weekly malaria incidence (the number of confirmed cases per week divided by the total population of the village) was determined for each of the 575 village. A square root transformation was applied to these incidence rates to stabilize the variance [4]. The functional data method [4, 5] states that the observed square root of incidence rate for a village i at a week j is the sum of a function on the week j and an error term:

$$y_{ij} = \sqrt{Inc_{ij}} = x_i(t_j) + \varepsilon_{ij} \quad \text{i} = 1, \ldots, 575, \quad \text{j} = 1, \ldots, 261 \quad (\text{eq. } 1)$$

where x_i is a regular (smooth) function which describes the temporal pattern of malaria incidence in village i, t_j is the (continuous) calendar time, and ε_{ij} is an error term representing the difference between the function value and the observed data for village i at week j.

The function x_i is defined as follows:

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\, \varphi_k\,(t) \quad (eq. 2)$$

where φ_k are basis functions, K represents the total number of basis functions and c_{ik} are the coefficients obtained by the least squares method by minimizing the penalized error sum of squares.

$$SSE(x_i) = \sum_{j=1}^{T} (y_{ij} - x_i(t_j))^{\,2} + \lambda \int_P \left| x_i''(t) \right|^2 dt \quad , i = 1, \ldots, 575 \quad (eq. 3)$$

where λ is the smoothing parameters and x_i'' the second derivative function with $\int_P |x_i''(t)|^2\, dt < \infty$.

To estimate the underlying function or functional data x_i, the family of basis functions φ_k, their total number K and the smoothing parameter λ should be chosen. The basis functions are families of known and mathematically independent functions [4, 5]. Several basis functions are possible but they have to be determined according to the nature of the data. In this work, we used cubic B-splines to avoid periodic smoothing [4, 5]. While the choice of the smoothing parameter is very important, there is no universal rule for an optimal choice. However, a number of criteria are available, including the generalized cross-validation (GCV), which we used [23].

-Dissimilarity measures and hierarchical ascending clustering on smooth functions

To perform a hierarchical ascending clustering on smooth functions (functional data), a proximity measure between them was necessary. Several dissimilarity measures are proposed for time series. We focused on those based on value and behaviour that would adapt to the functional data [23–27]. For those based only on values we had: the Euclidean distance in a multidimensional space (d_{EUC}), the Lp-metric estimating the surface between 2 functional data (d_{FDA}), the dynamic time warping dissimilarity measure (d_{DTW}) and the discrete wavelet transformation dissimilarity measure (d_{DWT}). To obtain dissimilarity measures based on values and behaviour (d_{CORT}), a temporal correlation between 2 functional data, were combined with each of the above dissimilarity measures based on values only [24, 26].

$$d_{CORT}(x_i, x_{i'}) = f_\xi[CORT(x_i, x_{i'})] * d(x_i, x_{i'})$$

$$CORT(x_i, x_{i'}) = \frac{\sum_{t=1}^{T-1}(x_i(\mathrm{t}+1) - x_i(\mathrm{t}))(x_{i'}(\mathrm{t}+1) - x_{i'}(\mathrm{t}))}{\sqrt{\sum_{t=1}^{T-1}(x_i(\mathrm{t}+1) - x_i(\mathrm{t}))^2}\sqrt{\sum_{t=1}^{T-1}(x_{i'}(\mathrm{t}+1) - x_{i'}(\mathrm{t}))^2}}$$

The adaptative function $f_\xi(u) = \frac{2}{1+\exp(\xi u)}$, $\xi \geq 0$ is used to adjust the percentage of contribution of value and behavior according to the value of the parameter ξ.

Table 1 here

Table 1: The percentage of contribution in d_{CORT} dissimilarity measure according to the parameter ξ.

ξ	Behavior Contribution (%)	Values Contribution (%)
0	0	100
1	46.2	53.7
2	76.2	23.8
3	90.5	9.4
≥ 5	~ 100	~ 0

151 At this stage only the Euclidean distance and the dynamic time warping distance were
152 implemented in R package with this dissimilarity measure. We have written a R program to
153 estimate the functional Euclidean distance and the discrete wavelet transformation dissimilarity
154 based on valued and behaviour with the same formula above. When ξ=0 (Table 1) we had the
155 4 dissimilarity measures based only on values, with the other values ξ , a total of 20 dissimilarity
156 measures were used in this analysis.

157 Thus, hierarchical ascending clustering (HAC) was performed on smooth functions with each
158 dissimilarity measure with Ward aggregation method [28]. The optimal number of clusters with
159 the best dissimilarity measure was guided by three validity indices used in a multidimensional
160 space for functional data [29–31]: Connectivity, Dunn, Silhouette and in addition the percentage
161 of inertia explained R2. We were looking for a classification (a result of HAC) with connectivity
162 index close to 0, high Dunn index, Silhouette and R2 close to 1. Thus, a Principal Component
163 Analysis PCA [32] was performed on these indices to choose the best classification.

164 In these two first step we worked on the functional data of the square root transformation of
165 observed time series, but for the following step we applied square transformations to have the
166 functional data corresponding to the observed time series.

167 After the clustering of the 575 functional data into Q clusters, we defined the functional data of
168 each cluster by $C_q(t)$, q=1, …, Q being the cumulative weekly incidence of the villages
169 belonging to the cluster q. Then, their 95% point-wise confidence intervals were computed by
170 adding and subtracting two of the standard errors, that is, the square root of the sampling
171 variances, to the actual fit [4].

172 ***-Velocity and acceleration (slope and rate of change of the slope)***

173 To further describe the malaria incidence pattern of each cluster, the first and second derivative
174 were determined for each functional data of a cluster. Thus, with mathematical properties of

univariate function optimization [33] and one-dimensional kinematics in physics [34], seven epidemiological indicators based on velocity and acceleration have been defined (Figure 1, Table 2). These epidemiological indicators were: The beginning of seasonal outbreaks and the start acceleration of the growth phase (A); The beginning of the pre-slowdown of the growth phase (B); The deceleration's beginning of growth phase (C); The peak (D) also corresponding after to the beginning of the acceleration of the decrease phase; The beginning of the deceleration of the decrease phase (E); The beginning of the tail (F); The end of the seasonal outbreaks (G). Finally, a Principal component analysis PCA was performed on the durations: AB, AD, CE, DG, FG, BF and AG for seasonal outbreaks of each cluster.

Table 2 here

Table 2

The description of epidemiological indicators and the determination of their corresponding date for a functional cluster C_q

Epidemiological indicators (EI)	Determination of EI's dates
Beginning of seasonal outbreaks and the start acceleration of the growth phase (A)	$t_A = \begin{cases} first\ t\ such\ {C_q}'(t) > 0 \\ {C_q}''(t) > 0 \\ on\ 3\ weeks \end{cases}$
Beginning of the pre-slowdown of the growth phase (B)	$t_B = \left\{ \underset{t\ such\ {C_q}'(t)>0}{\operatorname{argmax}} \left({C_q}''(t)\right) \right.$
deceleration's beginning of growth phase (C)	$t_C = \left\{ \underset{t\ such\ {C_q}''(t)=0}{\operatorname{argmax}} \left({C_q}'(t)\right) \right.$
Peak of seasonal outbreaks and beginning of the acceleration of the decrease phase (D)	$t_D = \begin{cases} {C_q}'(t) = 0 \\ {C_q}''(t) < 0 \end{cases}$
Beginning of the deceleration of the decrease phase (E)	$t_E = \left\{ \underset{t\ such\ {C_q}''(t)=0}{\operatorname{argmin}} \left({C_q}'(t)\right) \right.$
Beginning of the tail of seasonal outbreaks (F)	$t_F = \left\{ \underset{t\ such\ {C_q}'(t)<0}{\operatorname{argmax}} \left({C_q}''(t)\right) \right.$
End of seasonal outbreaks (G)	$t_G = \begin{cases} first\ t\ such\ {C_q}'(t) = 0 \\ on\ 3\ weeks \end{cases}$

Figure 1 here

Figure 1

A graphical example for the seven epidemiological indicators: the beginning of seasonal outbreaks and the start acceleration of the growth phase (A); the beginning of the pre-slowdown of the growth phase (B); the deceleration's beginning of growth phase (C); the peak (D) also corresponding after to the beginning of the acceleration of the decrease phase; the beginning of the deceleration of the decrease phase (E); the beginning of the tail (F); the end of the seasonal outbreaks (G); functional incidence in red line, functional velocity in black bold line (first derivative) and functional acceleration in black discontinuous line (second derivative).

Results

-From observed to smoothed malaria incidence

The observed incidence ranged from 0 to 16,667 cases/100,000 person-weeks (Figure 2, Panel A) at the village level. The maximum seasonal outbreaks peaks ranged from 4,478 and 16,667 cases/100,000 person-weeks. The search for the optimal number of basis function and the optimal smoothing parameter gave $K_{opt} = 110$ and $\lambda_{opt} = 103$ which minimized the error by GCV equal to 11.8 with a standard deviation of $\sigma = 0.12$. Smoothed malaria incidence ranged from 0 to 2,296 cases/100,000 person-weeks with seasonal outbreaks peaks between 541 and 2,296 cases/100,000 person-weeks (Figure 2, Panel B).

Figure 2 here

Figure 2

Weekly malaria incidence evolution for each village from January 2008 to December 2012: observed time series (Panel A) and smoothed functions of time series (Panel B)

-Identification of malaria incidence patterns

Three clusters with the DTWCORT1 dissimilarity measure were chosen based on the principal component analysis performed on validity indices and dissimilarity measures for 3 and 4 number classes (Additional file 1, Additional file 2).

Figure 3 here

Figure 3

Principal component analysis on validity indices and dissimilarity measures for 3 and 4 number clusters: validity indices map (Variables, Panel A), dissimilarity measures map (Individuals, Panel B)

Indeed, the dimension 1 represented high Dunn index and silhouettes, and low connectivity (Figure 3, Panel A). The dimension 2 essentially represented the percentage of inertia explained by the clusters (Figure 3, Panel A). The best classification should therefore be located in the upper right of the factorial plane of the dissimilarity measures and the number of clusters (Figure 3, Panel B). The DTWCORT1 dissimilarity measure took account 46.2% of the temporal correlation between functional data and 53.7% of the geometric distance.

Figure 4 here

Figure 4

The spatial distribution of malaria incidence patterns in the area study (Panel A) and smoothed functions for each malaria incidence pattern (Panel B): Senegal map and the location of the study area pointed by the arrow, high incidence pattern in red line, intermediate incidence pattern in blue line and low incidence pattern in green line.

The high incidence pattern (high pattern) consisted of a set of 12 villages with the highest observed average incidence over the 5 year study period (114 cases/1000 person-years), mainly located in the southern part of the study area (Figure 4, Panel A). Its smoothed seasonal outbreaks peaks ranged from 233 to 884 cases/100,000 person-weeks (Table 3).

The intermediate incidence pattern (intermediate pattern) included 97 villages had 13 cases/1000 person-years as observed average incidence over the study period, located both in the southern and northern part of the study area. Its smoothed seasonal outbreaks peaks ranged from 26 to 131 cases/100,000 person-weeks (Table 3).

The low incidence pattern (low pattern) consisted of a set of 466 villages with the lowest average incidence over study period (2.6 cases/1000 person-years), mainly located in the northern part of the study area. Its smoothed seasonal outbreaks peaks ranged from 7 to 34 cases/100,000 person-weeks (Table 3).

The two higher incidence patterns (high and intermediate) correspond to 23% of the population and 19% of villages.

Table 3 here

Table 3

Incidence description of Malaria incidence patterns: the type of pattern, their number villages and their ranges peaks of smoothed seasonal outbreaks and their observed cumulative incidence over the 5 years of study period

Malaria incidence Patterns	Number of villages	Range peaks of smoothed seasonal outbreaks (cases/100,000 person-weeks)	Observed cumulative incidence over the 5 year study period (cases/1000 person_years)
High	12	233-884	114
Intermediate	97	26-131	13
Low	466	7-34	2.6

Figure 5 here

Figure 5

Weekly observed malaria incidence in black solid line, smoothed malaria incidence in color solid line and smooth 95% point-wise confidence intervals in discontinuous color line : high incidence pattern in red (Panel A), intermediate incidence pattern in blue (Panel B) and low incidence pattern in green (Panel C).

The observed incidence of the patterns, their smoothed incidence and their 95% point-wise confidence intervals of smoothing were highlighted for each malaria incidence pattern (Figure 5). In all the patterns, the observed incidence rates were in the ranges except for a few peaks in the high pattern (Figure 5).

-Velocity and acceleration of malaria incidence patterns

The velocity curve of each pattern (Additional file 3, Panel A) represented the variation over time of the slope and the acceleration curve of each pattern (Additional file 3, Panel B) represented the rate evolution of change of the slope. The velocities and accelerations of the high pattern were higher, followed by those of the intermediate pattern and those of the low pattern were the lowest (Additional file 3).

The epidemiological indicators determined using the velocity and acceleration functions (Figure 6) as described in the methodology showed that the low pattern was always the one that started (A) the latest. The high pattern started earlier 3 times during the 5 seasonal outbreaks and the intermediate pattern twice. In addition, seasonal outbreaks of the high and intermediate patterns usually started between April and June with a lag between 1 to 3 weeks. Those of the low pattern started between June and July with a delay between 4 and 9 weeks after the intermediate pattern, and with a lag between 3 and 10 weeks after the high pattern (Table 4).

The phases of pre-slowdown (B) and slowdown (C) of epidemic's growth started mainly between August and September for all patterns with a lag between 1 and 2 weeks. Then, the peak (D) of seasonal outbreaks for all patterns, occurred between October and November almost the same time or with maximum 1 week of lag. The beginning of the deceleration phase of the decrease (E) occurred between November and December for all patterns, almost at the same time or with maximum 1 week of lag. The exception of the latter point was that the E of seasonal

outbreaks beginning in 2009 and 2010 of the high pattern and those beginning in 2010 of the intermediate pattern began between January and February of their following years respectively (Table 4).

The tails (F) of seasonal outbreaks for low pattern were the earliest, starting in December; those of high pattern were the latest, starting between December and March. Those of intermediate pattern followed those of high pattern and started between December and February. Moreover, the lag between high and low pattern was from 1 to 11 weeks, those between high and intermediate pattern was from 1 to 9 weeks and those between intermediate and low pattern was from 0 to 7 weeks (Table 4).

The end of seasonal outbreaks (G) for the high and intermediate pattern occurred between March and May with a lag from 0 to 7 weeks; those of low pattern occurred the earliest between February and March with a lag from 3 to 13 weeks before high pattern and a lag from 3 to 9 weeks before the intermediate pattern.

Figure 6 here

Figure 6

Smoothed incidence in color solid line, their velocity in black bold solid line, their acceleration in black discontinuous line and the epidemiological indicator of their seasonal outbreaks (A: onset, B: near slowdown of growth, C: beginning slowdown of growth, D: Peak, E: beginning acceleration of decline, F: beginning of tail , G: end): high incidence pattern in red (Panel A), intermediate incidence pattern in blue (Panel B) and low incidence pattern in green (Panel C).

Table 4 here

312 **Table 4** : The epidemiological indicators (EI) and their characteristics over seasonal outbreaks

Start year seasonal outbreak	EI	DateHigh	DateInter	DateLow	WeekHigh	WeekInter	WeekLow
2008	A	13/05/2008	29/04/2008	17/06/2008	20	18	25
2009	A	19/05/2009	26/05/2009	28/07/2009	21	22	31
2010	A	08/06/2010	01/06/2010	29/06/2010	24	23	27
2011	A	26/04/2011	10/05/2011	14/06/2011	18	20	25
2012	A	03/04/2012	24/04/2012	29/05/2012	15	18	23
2008	B	09/09/2008	02/09/2008	26/08/2008	37	36	35
2009	B	25/08/2009	08/09/2009	25/08/2009	35	37	35
2010	B	14/09/2010	31/08/2010	07/09/2010	38	36	37
2011	B	23/08/2011	23/08/2011	23/08/2011	35	35	35
2012	B	28/08/2012	28/08/2012	28/08/2012	36	36	36
2008	C	30/09/2008	23/09/2008	23/09/2008	40	39	39
2009	C	22/09/2009	22/09/2009	15/09/2009	39	39	38
2010	C	05/10/2010	28/09/2010	28/09/2010	41	40	40
2011	C	13/09/2011	20/09/2011	20/09/2011	38	39	39
2012	C	18/09/2012	18/09/2012	25/09/2012	39	39	40
2008	D	28/10/2008	21/10/2008	21/10/2008	44	43	43
2009	D	27/10/2009	20/10/2009	20/10/2009	44	43	43
2010	D	02/11/2010	26/10/2010	02/11/2010	45	44	45
2011	D	11/10/2011	25/10/2011	18/10/2011	42	44	43
2012	D	23/10/2012	23/10/2012	30/10/2012	44	44	45
2008	E	02/12/2008	02/12/2008	25/11/2008	49	49	48
2009	E	16/02/2010	01/12/2009	08/12/2009	8	49	50
2010	E	18/01/2011	18/01/2011	30/11/2010	4	4	49
2011	E	29/11/2011	29/11/2011	22/11/2011	49	49	48
2012	E	27/11/2012	20/11/2012	27/11/2012	49	48	49
2008	F	06/01/2009	23/12/2008	23/12/2008	2	52	52
2009	F	16/03/2010	12/01/2010	29/12/2009	12	3	1
2010	F	15/02/2011	08/02/2011	21/12/2010	8	7	52
2011	F	20/12/2011	13/12/2011	13/12/2011	52	51	51
2008	G	12/05/2009	24/03/2009	10/02/2009	20	13	7
2009	G	11/05/2010	04/05/2010	02/03/2010	20	19	10
2010	G	26/04/2011	26/04/2011	22/03/2011	18	18	13
2011	G	20/03/2012	03/04/2012	28/02/2012	13	15	10

313

314 The seasonal outbreaks for all patterns were further described with the PCA performed on the
315 durations between selected relevant epidemiological indicators (Figure 7, Panel A). These were:
316 the duration of growth's acceleration phase (AB); those between start and peak (AD); those
317 between slowdown of growth and decline (CE) indicating the width of the peak area; those

between peak and the end (DG); those between the tail and the end of seasonal outbreaks (FG); those between pre-slowdown and the tail (BF) indicating the width of epidemic's body; those between the start and the end of epidemic episodes (AG) indicating the duration of the seasonal outbreaks.

Figure 7 here

Figure 7

Principal component analysis on duration epidemiological indicator and seasonal outbreaks of the patterns: epidemiological indicator map (Variables, Panel A), seasonal outbreaks of the patterns map (Individuals, Panel B)

The result of PCA (Figure 7, Additional file 4) showed that the seasonal outbreaks (Figure 7, Panel B) of high pattern starting since 2009 (2009H) and 2010 (2010H), and those of the intermediate pattern starting since 2010 (2010I) were mainly characterized by a high BF and CE, and also by a low FG.

In addition, the seasonal outbreaks of low pattern were characterized by low AG, DG AD and AB. The seasonal outbreaks starting since 2008 and 2011 for high and intermediate patterns (2008H, 2008I, 2011I, and 2011H) were mainly characterized by high FG on the one hand and on the other hand by low BF and CE. In addition, 2008H, 2009I and 2011H were also characterized by a high AG, AD, AB, and DG.

Discussion

The approach used here led to the identification of three distinct patterns for the time-course of malaria incidence in a village, by taking into account dynamics of malaria incidence over the whole study period.

342 In the estimation of the smooth functions (functional data) we used B-Spline cubic basis
343 function. Indeed, some other basis functions were available as the Fourier basis for periodic
344 data. We used B-spline cubic basis, because, even if seasonal outbreaks occurred with a
345 periodicity, intensities were different. However, in our context, there are no difference between
346 both basis, and, in this case, B-spline are more tractable [4, 5, 7].

347 The choice of dissimilarity measure for functional data is important before applying an
348 unsupervised classification method, to have well-separated classes. Some other dissimilarity
349 measures could be added [35]. We preferred to limit them on the measures less dependent to
350 the autocorrelation structure. Indeed, the smoothing approach of functional data may impact
351 the autocorrelation structure. For the choice of validity indices, we preferred also to concentrate
352 on a small number of those assessing the separability (Dunn), compactness (Connectivity), the
353 quality of clustering for villages in average (Silhouette) [30, 31] and percentage of inertia (R^2).

354 The detection methods of transmission foci or hotspots have been defined differently in the
355 literature[36]. There are methods that define them from an incidence or prevalence threshold
356 [14] others with biological parameter [37] and others from scanning algorithm [9] or
357 geostatistical approaches [8] . In addition, spatial and temporal analyzes were often based on
358 the fragmentation of the study period. Indeed in some researches, these temporal divisions were
359 based on the calendar (month, year) or the rainy seasons , in other works of temporal
360 fragmentation methods were based on algorithms such as change point analysis [14–18, 38, 39].

361 In our study, hotspot identification was made by taking account not only the value of the
362 incidence but also the dynamic of the malaria incidence over the whole study period hence the
363 malaria incidence pattern name. Consequently, this method can be used to distinguish two
364 spatial units that have the same level of incidence or the same number of cases, but with
365 different dynamics. Indeed, an epidemic that starts with a high intensity and declines over time
366 is different from another that increases over time, leading to different control strategies.

With our approach, characterizations of the seasonal outbreaks have been made using the velocities and accelerations of the malaria incidence pattern. This allowed us to define epidemiological indicators for which seasonal outbreaks were further described. Following this approach, one of the results showed that low incidence pattern was the latest to start and the earliest to end seasonal outbreaks, all incidence pattern reached their seasonal peaks almost in the same time. In the case of other countries, different results can be found with these epidemiological indicators. In addition, the seasonal outbreaks 2009H, 2010H and 2010I were remarkable. The focus analyzed on their velocity and acceleration showed an outbreak rebound. This can have impact on their high width of peak area (CE) and body (BF). Despite this, their durations tail were low, consequently the duration of the outbreak as well.

Furthermore, previous researches had focused on the search for epidemic thresholds and stratification into intensity levels of different epidemics, particularly in the field of influenza surveillance and acute respiratory infection in Europe [40–44]. However, as stated by numerous authors, there was no automatic and objective way to compare thresholds and intensity levels across the countries studied. Although the epidemiological contexts are not the same with malaria, we were able to introduce an approach based on functional data allowing the smoothing of the time series of the village incidence by a single smoothing parameter and a single number of basis functions. Thereby, this allowed a possible comparison between them because they had the same scale [4]. For this purpose, even if this was not our main objective, we could define the starting date of an outbreak as the time from which the velocity and acceleration functions are strictly positive for at least 3 consecutive weeks (indeed, the first symptoms of malaria appear 1 to 4 weeks after infection [45]). Thus, this approach can be applied in other diseases contexts.

This work showed a relatively small number of high incidence villages, which were adjacent to low incidence villages. It may be useful to investigate social and environmental factors that

may be associated with locally high incidence (e.g. proximity to water bodies, use of control measures,…). The two higher incidence patterns correspond to 23% of the population. Awareness of these trends may assist district health teams to strengthen control in high risk communities and guide targeted intervention, and our results suggest that a targeted strategy might need to include about 20% of the population.

Conclusion

Functional data analysis can be used to classify communities based on time series of malaria incidence, and to identify high incidence communities. Indicators can be derived from the fitted functions which characterize the timing of outbreaks. These tools may help to better target control measures.

Ethics

The protocols for the field studies were approved by Senegal's Conseil National pour la Recherche en Santé and the ethics committee of the London School of Hygiene & Tropical Medicine. The SMC trial were registered at www.clinicaltrials.gov, number NCT 00712374.

Consent for publication

All authors have read and approved the manuscript.

Availability of data and materials

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable in our study.

Authors' contributions

SD and JG designed the study, performed data processing, the statistical analysis and interpretation, and wrote the first draft of the article; PiM et AG contributed to the statistical analysis; KS contributed to the data processing; EB, BC, CS and PaM coordinated the data collection and validation; PC, PaM contributed to the interpretation of the results. All authors read and approved the final manuscript.

Acknowledgments

We thank the public health network “Reseau doctoral en santé publique” coordinated by the EHESP (School for Higher Studies in Public Health) for supporting the thesis project of SD. We thank the NGO PROSPECTIVE & COOPERATION for collaboration. Lastly, we thank Arianne Dorval for comments.

References

1. World Health Organization, Global Malaria Programme, World Health Organization. A framework for malaria elimination. 2017. http://apps.who.int/iris/bitstream/10665/254761/1/9789241511988-eng.pdf. Accessed 28 Oct 2019.

2. Ferraty F, Vieu P. Richesse et complexité des données fonctionnelles. Revue Modulad. 2011;43:25–43. http://www.modulad.fr/archives/numero-43/VIEU/2-Vieu.pdf.

3. Ferraty F. Modélisation statistique pour variables aléatoires fonctionnelles: théorie et application. Habilitationa diriger des recherches, Université Paul Sabatier. 2003. https://www.math.univ-toulouse.fr/~besse/pub/chapBC.ps.

4. Ramsay JO, Silverman BW. Functional data analysis. 2nd ed. New York: Springer; 2005.

5. Ramsay JO, Hooker G, Graves S. Functional data analysis with R and MATLAB. Springer Science & Business Media; 2009.

6. Ramsay JO, Silverman BW, Ramsay JO, Silverman BW. Applied Functional Data Analysis: Methods and Case Studies. Springer; 2002.

7. Ullah S, Finch CF. Applications of functional data analysis: A systematic review. BMC Med Res Methodol. 2013;13:43. doi:10.1186/1471-2288-13-43.

8. Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. :52.

9. Kulldorff M. A spatial scan statistic. Communications in Statistics - Theory and Methods. 1997;26:1481–96. doi:10.1080/03610929708831995.

10. Gaudart J, Graffeo N, Coulibaly D, Barbet G, Rebaudet S, Dessay N, et al. **SPODT** : An *R* Package to Perform Spatial Partitioning. Journal of Statistical Software. 2015;63. doi:10.18637/jss.v063.i16.

11. Bejon P, Williams TN, Nyundo C, Hay SI, Benz D, Gething PW, et al. A micro-epidemiological analysis of febrile malaria in Coastal Kenya showing hotspots within hotspots. eLife. 2014;3:e02130. doi:10.7554/eLife.02130.

12. Platt A, Obala AA, MacIntyre C, Otsyula B, Meara WPO. Dynamic malaria hotspots in an open cohort in western Kenya. Scientific Reports. 2018;8:647. doi:10.1038/s41598-017-13801-6.

13. Sallah K, Giorgi R, Ba EH, Piarroux M, Piarroux R, Griffiths K, et al. Targeting hotspots to reduce transmission of malaria in Senegal: modeling of the effects of human mobility. 2018. doi:10.1101/403626.

14. Landier J, Parker DM, Thu AM, Lwin KM, Delmas G, Nosten FH, et al. Effect of generalised access to early diagnosis and treatment and targeted mass drug administration on Plasmodium falciparum malaria in Eastern Myanmar: an observational study of a regional elimination programme. The Lancet. 2018;0. doi:10.1016/S0140-6736(18)30792-X.

15. Bejon P, Williams TN, Liljander A, Noor AM, Wambua J, Ogada E, et al. Stable and Unstable Malaria Hotspots in Longitudinal Cohort Studies in Kenya. PLoS Med. 2010;7. doi:10.1371/journal.pmed.1000304.

16. Coulibaly D, Travassos MA, Tolo Y, Laurens MB, Kone AK, Traore K, et al. Spatio-Temporal Dynamics of Asymptomatic Malaria: Bridging the Gap Between Annual Malaria Resurgences in a Sahelian Environment. The American Journal of Tropical Medicine and Hygiene. 2017;97:1761–9. doi:10.4269/ajtmh.17-0074.

17. Ouedraogo B, Inoue Y, Kambiré A, Sallah K, Dieng S, Tine R, et al. Spatio-temporal dynamic of malaria in Ouagadougou, Burkina Faso, 2011–2015. Malaria Journal. 2018;17. doi:10.1186/s12936-018-2280-y.

18. Sissoko MS, Sissoko K, Kamate B, Samake Y, Goita S, Dabo A, et al. Temporal dynamic of malaria in a suburban area along the Niger River. Malar J. 2017;16. doi:10.1186/s12936-017-2068-5.

19. Santos-Vega M, Bouma MJ, Kohli V, Pascual M. Population Density, Climate Variables and Poverty Synergistically Structure Spatial Risk in Urban Malaria in India. PLOS Neglected Tropical Diseases. 2016;10:e0005155. doi:10.1371/journal.pntd.0005155.

20. Cissé B, Ba EH, Sokhna C, NDiaye JL, Gomis JF, Dial Y, et al. Effectiveness of Seasonal Malaria Chemoprevention in Children under Ten Years of Age in Senegal: A Stepped-Wedge Cluster-Randomised Trial. PLOS Medicine. 2016;13:e1002175. doi:10.1371/journal.pmed.1002175.

21. Bâ E-H, Pitt C, Dial Y, Faye SL, Cairns M, Faye E, et al. Implementation, coverage and equity of large-scale door-to-door delivery of Seasonal Malaria Chemoprevention (SMC) to children under 10 in Senegal. Sci Rep. 2018;8. doi:10.1038/s41598-018-23878-2.

22. Bulletin Epidemiologique ANNUEL 2018 du Paludisme au SENEGAL (www.pnlp.sn).

23. Febrero-Bande M, de la Fuente MO, others. Statistical computing in functional data analysis: The R package fda. usc. Journal of Statistical Software. 2012;51:1–28. https://www.jstatsoft.org/htaccess.php?volume=51&type=i&issue=04&paper=true. Accessed 13 Sep 2017.

24. Douzal-Chouakria A, Amblard C. Classification trees for time series. Pattern Recognition. 2012;45:1076–91. doi:10.1016/j.patcog.2011.08.018.

25. Giorgino T. Computing and Visualizing Dynamic Time Warping Alignments in *R* : The **dtw** Package. J Stat Soft. 2009;31. doi:10.18637/jss.v031.i07.

26. Chouakria AD, Nagabhushan PN. Adaptive dissimilarity index for measuring time series proximity. ADAC. 2007;1:5–21. doi:10.1007/s11634-006-0004-6.

27. Montero P, Vilar JA. **TSclust** : An *R* Package for Time Series Clustering. J Stat Soft. 2014;62. doi:10.18637/jss.v062.i01.

28. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif. 2014;31:274–95. doi:10.1007/s00357-014-9161-z.

29. Dunn† JC. Well-Separated Clusters and Optimal Fuzzy Partitions. Journal of Cybernetics. 1974;4:95–104. doi:10.1080/01969727408546059.

30. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53–65. doi:10.1016/0377-0427(87)90125-7.

31. Malouche D. Méthodes de classifications. 2013. http://math.univ-bpclermont.fr/DoWellB/docs/malouche/methodes_classifications_CF_Juin2013.pdf.

32. Husson F, Lê S, Pagès J. Exploratory multivariate analysis by example using R. Boca Raton: CRC Press; 2011.

33. Guichard D. 5. Curve Sketching. https://www.whitman.edu/mathematics/calculus_online/chapter05.html. Accessed 19 Nov 2019.

516 34. Sunil Kumar Singh. Acceleration and deceleration - Kinematics fundamentals - OpenStax
517 CNX. 2010. http://cnx.org/contents/f25d0bfc-5f61-411b-bcee-be8187ad5cc7@10. Accessed
518 18 Nov 2019.

519 35. Montero P, Vilar JA. Tsclust: An r package for time series clustering. Journal of. 2014.

520 36. McKay HS, Lessler J, Moore SM, Azman AS. What is a Hotspot Anyway? The American
521 Journal of Tropical Medicine and Hygiene. 2017;96:1270–3. doi:10.4269/ajtmh.16-0427.

522 37. Bousema T, Griffin JT, Sauerwein RW, Smith DL, Churcher TS, Takken W, et al. Hitting
523 Hotspots: Spatial Targeting of Malaria for Control and Elimination. PLoS Medicine.
524 2012;9:e1001165. doi:10.1371/journal.pmed.1001165.

525 38. Gaudart J, Poudiougou B, Dicko A, Ranque S, Toure O, Sagara I, et al. Space-time
526 clustering of childhood malaria at the household level: a dynamic cohort in a Mali village.
527 BMC Public Health. 2006;6. doi:10.1186/1471-2458-6-286.

528 39. Rouamba T, Nakanabo-Diallo S, Derra K, Rouamba E, Kazienga A, Inoue Y, et al.
529 Socioeconomic and environmental factors associated with malaria hotspots in the Nanoro
530 demographic surveillance area, Burkina Faso. BMC Public Health. 2019;19.
531 doi:10.1186/s12889-019-6565-z.

532 40. Fleming DM, Zambon M, Bartelds AIM, de Jong JC. The duration and magnitude of
533 influenza epidemics: A study of surveillance data from sentinel general practices in England,
534 Wales and the Netherlands. 1999;:7.

535 41. Rakocevic B, Grgurevic A, Trajkovic G, Mugosa B, Sipetic Grujicic S, Medenica S, et al.
536 Influenza surveillance: determining the epidemic threshold for influenza by using the Moving
537 Epidemic Method (MEM), Montenegro, 2010/11 to 2017/18 influenza seasons. Euro Surveill.
538 2019;24. doi:10.2807/1560-7917.ES.2019.24.12.1800042.

539 42. Teklehaimanot HD, Schwartz J, Teklehaimanot A, Lipsitch M. Alert Threshold
540 Algorithms and Malaria Epidemic Detection. Emerg Infect Dis. 2004;10:1220–6.
541 doi:10.3201/eid1007.030722.

542 43. Vega T, Lozano JE, Meerhoff T, Snacken R, Beauté J, Jorgensen P, et al. Influenza
543 surveillance in Europe: comparing intensity levels calculated using the moving epidemic
544 method. Influenza and Other Respiratory Viruses. 2015;9:234–46. doi:10.1111/irv.12330.

545 44. Vega T, Lozano JE, Meerhoff T, Snacken R, Mott J, Lejarazu RO de, et al. Influenza
546 surveillance in Europe: establishing epidemic thresholds by the Moving Epidemic Method.
547 Influenza and Other Respiratory Viruses. 2013;7:546–58. doi:10.1111/j.1750-
548 2659.2012.00422.x.

549 45. Bartoloni A, Zammarchi L. Clinical Aspects of Uncomplicated and Severe Malaria.
550 Mediterr J Hematol Infect Dis. 2012;4. doi:10.4084/MJHID.2012.026.

551

552

Additional file

Additional file 1.

Additional file 1.docx

Validity indices performed on each hierarchical ascending clustering's results for 3 and 4 clusters: Connectivity, Dunn, Silhouette and the percentage of inertia explained R2

Additional file 2.

Additional file 2.jpeg

Dendrogram resulting of hierarchical clustering on smooth function with DTWCORT1 dissimilarity measure: 12 villages with high incidence pattern (red), 97 villages with intermediate incidence pattern (blue border) and 466 with a low incidence pattern (green border)

Additional file 3.

Additional file 3.jpeg

The velocity (Panel A) and the acceleration (Panel B) dynamics of malaria incidence patterns: high incidence pattern in red line, intermediate incidence pattern in blue line and low incidence pattern in green line.

Additional file 4.

Additional file 4.docx

The PCA results on duration epidemiological indicator (EI) and seasonal outbreaks of the patterns: the PCA indicator (correlation between EI and dimensions, cos2 measuring the quality of projection of EI in the map factor, the percentage of contribution and the coordinates of seasonal outbreaks); dimensions or axis of PCA.

Figures

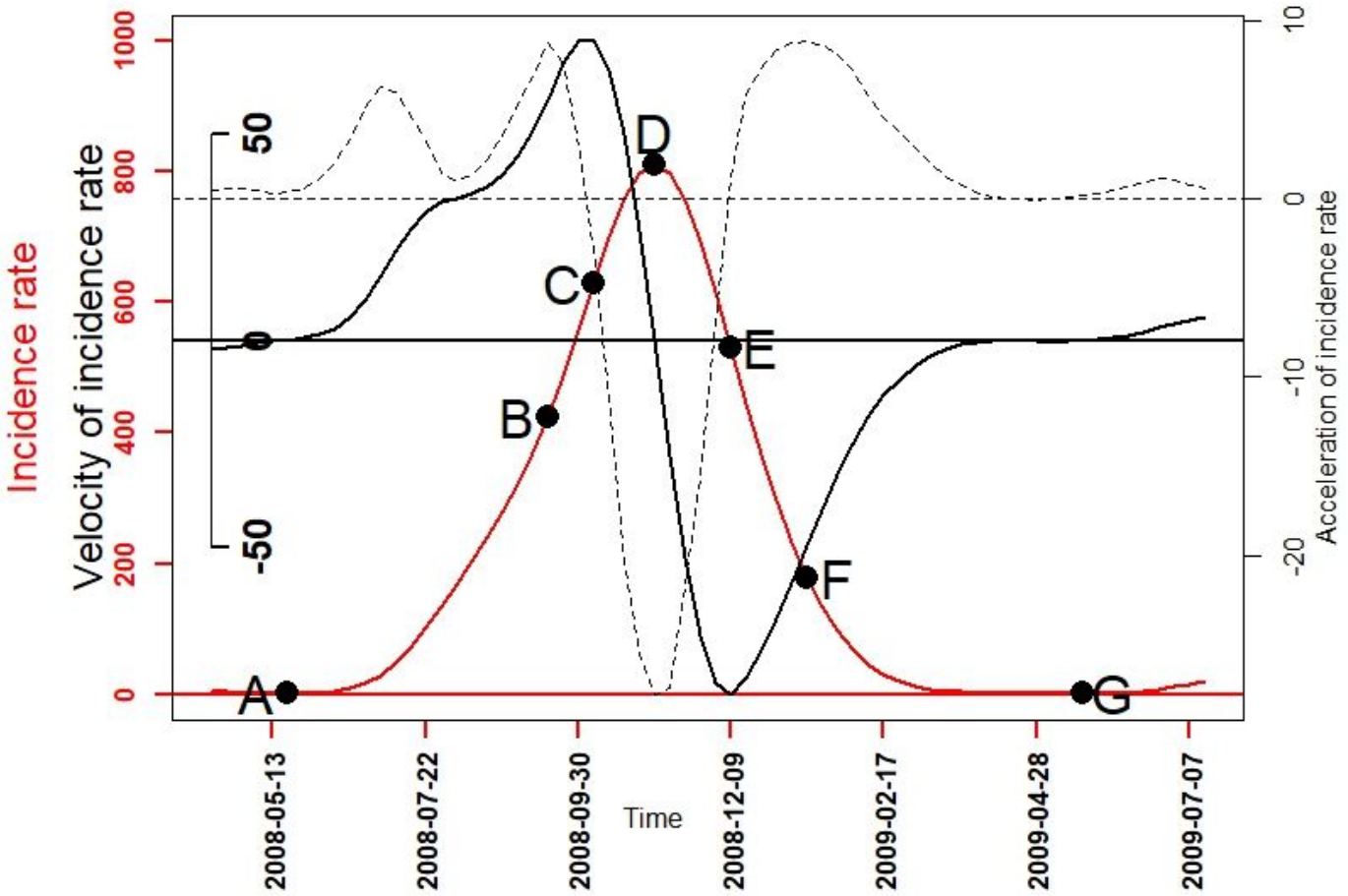

Figure 1

A graphical example for the seven epidemiological indicators: the beginning of seasonal outbreaks and the start acceleration of the growth phase (A); the beginning of the pre-slowdown of the growth phase (B); the deceleration's beginning of growth phase (C); the peak (D) also corresponding after to the beginning of the acceleration of the decrease phase; the beginning of the deceleration of the decrease phase (E); the beginning of the tail (F); the end of the seasonal outbreaks (G); functional incidence in red line, functional velocity in black bold line (first derivative) and functional acceleration in black discontinuous line (second derivative).

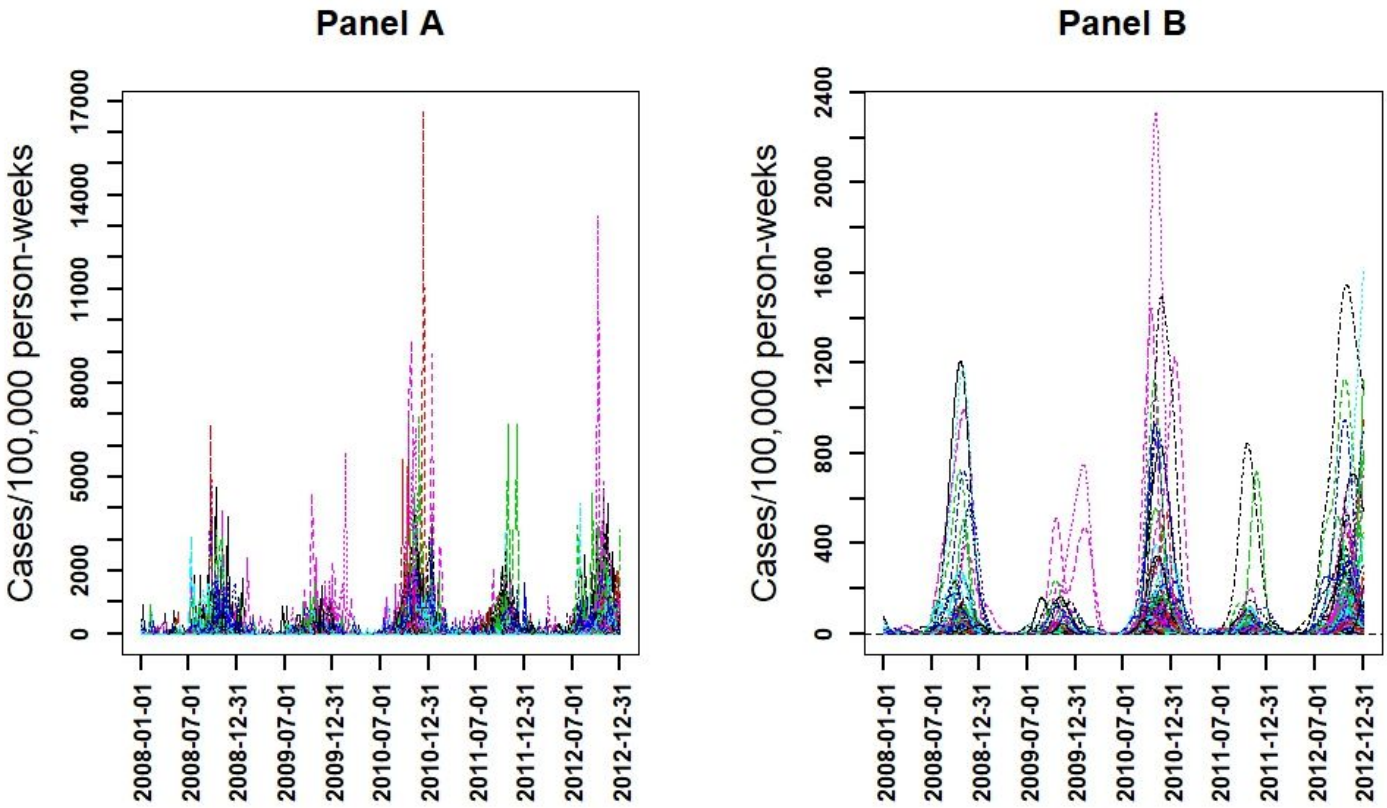

Figure 2

Weekly malaria incidence evolution for each village from January 2008 to December 2012: observed time series (Panel A) and smoothed functions of time series (Panel B)

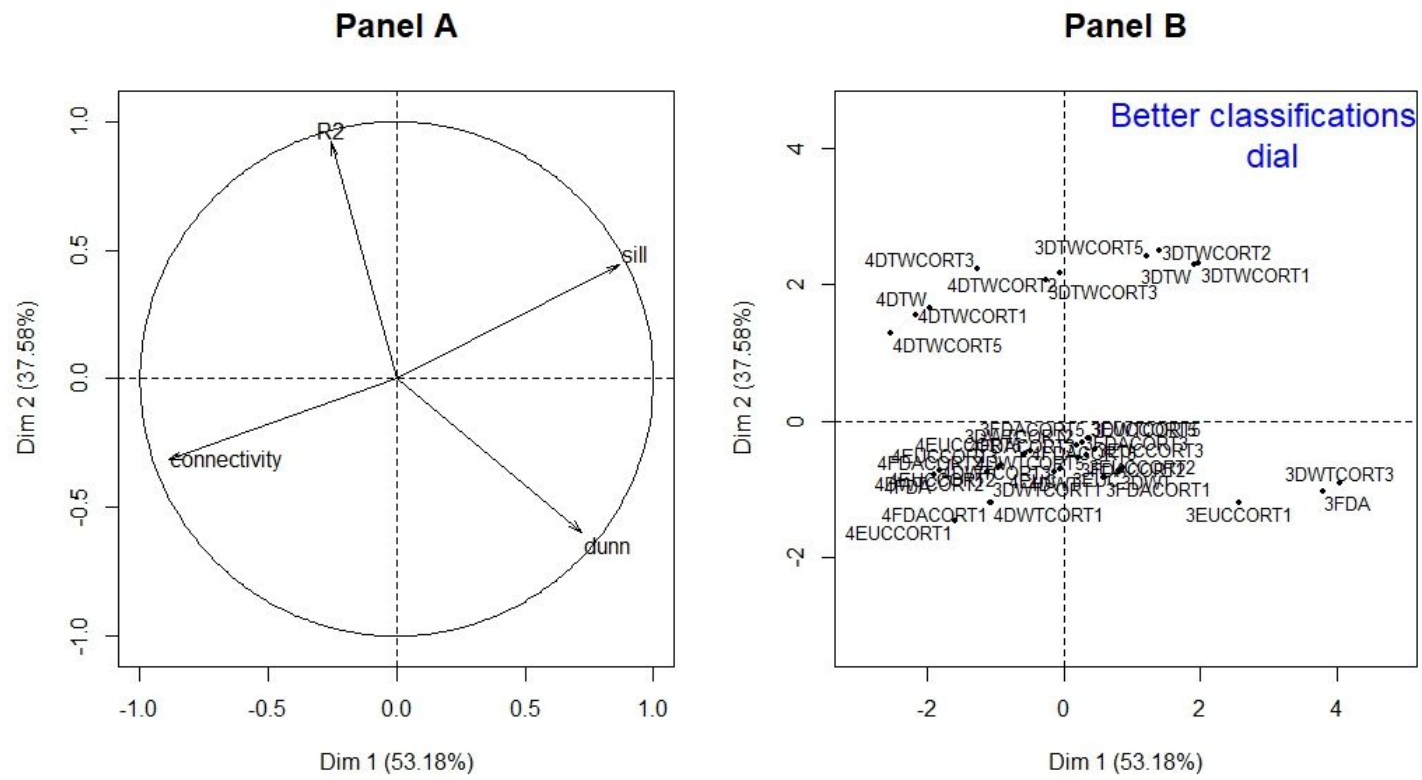

Figure 3

Principal component analysis on validity indices and dissimilarity measures for 3 and 4 number clusters: validity indices map (Variables, Panel A), dissimilarity measures map (Individuals, Panel B)

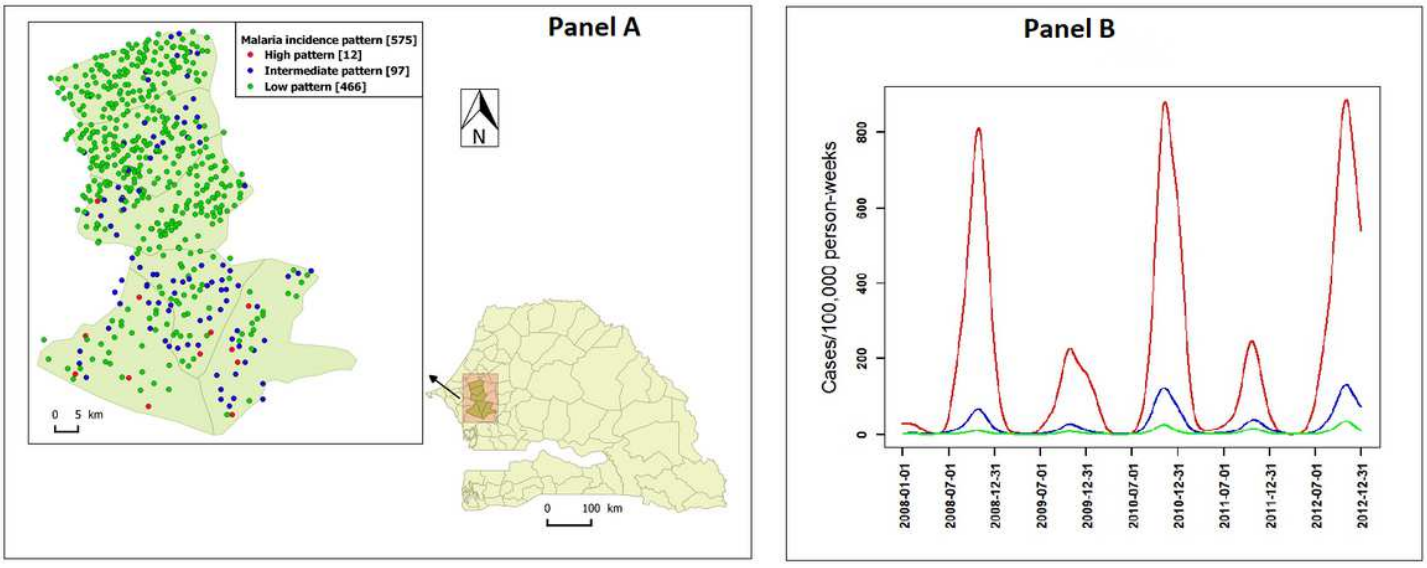

Figure 4

The spatial distribution of malaria incidence patterns in the area study (Panel A) and smoothed functions for each malaria incidence pattern (Panel B): Senegal map and the location of the study area pointed by

the arrow, high incidence pattern in red line, intermediate incidence pattern in blue line and low incidence pattern in green line.

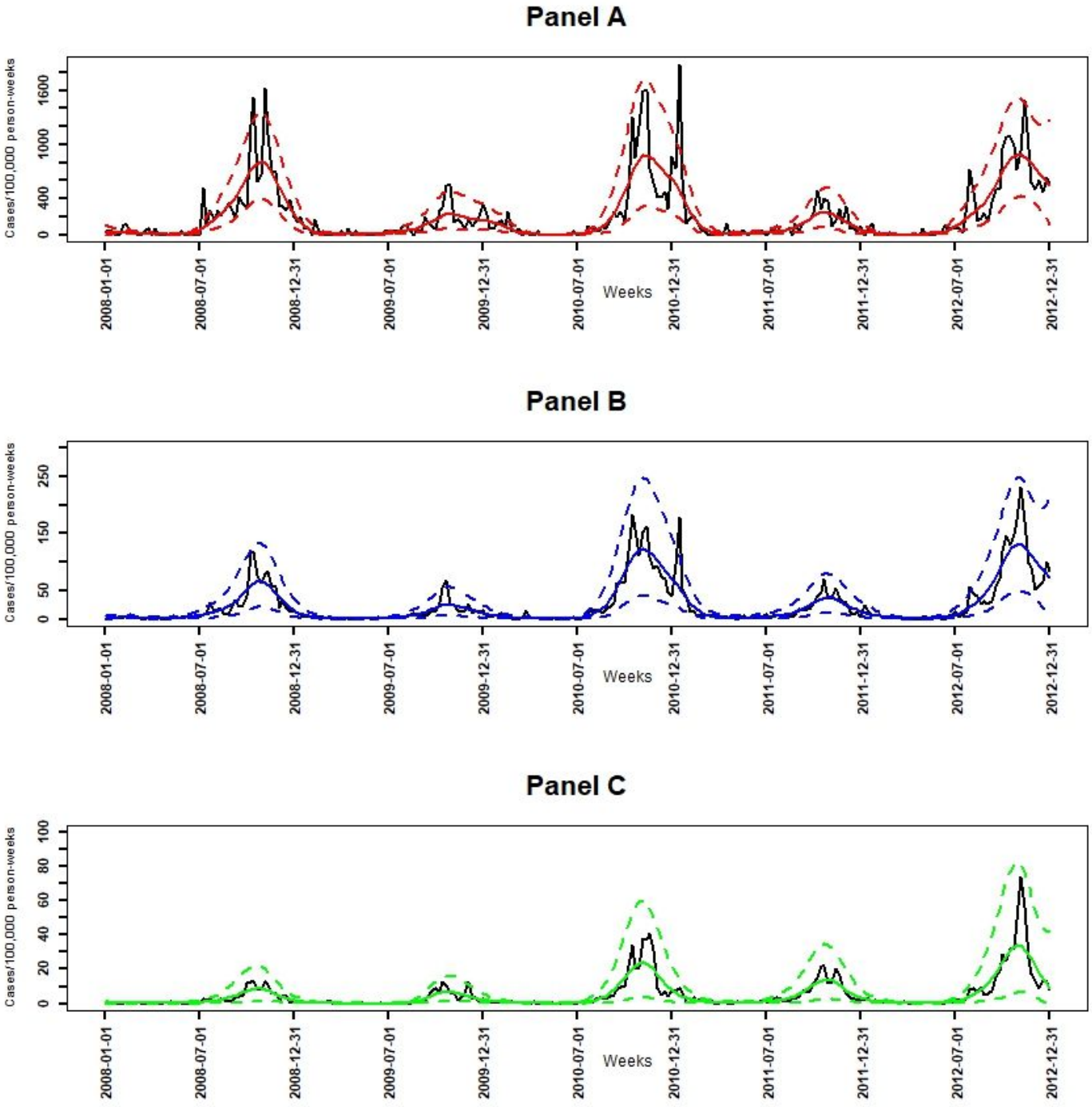

Figure 5

Weekly observed malaria incidence in black solid line, smoothed malaria incidence in color solid line and smooth 95% point-wise confidence intervals in discontinuous color line : high incidence pattern in red (Panel A), intermediate incidence pattern in blue (Panel B) and low incidence pattern in green (Panel C).

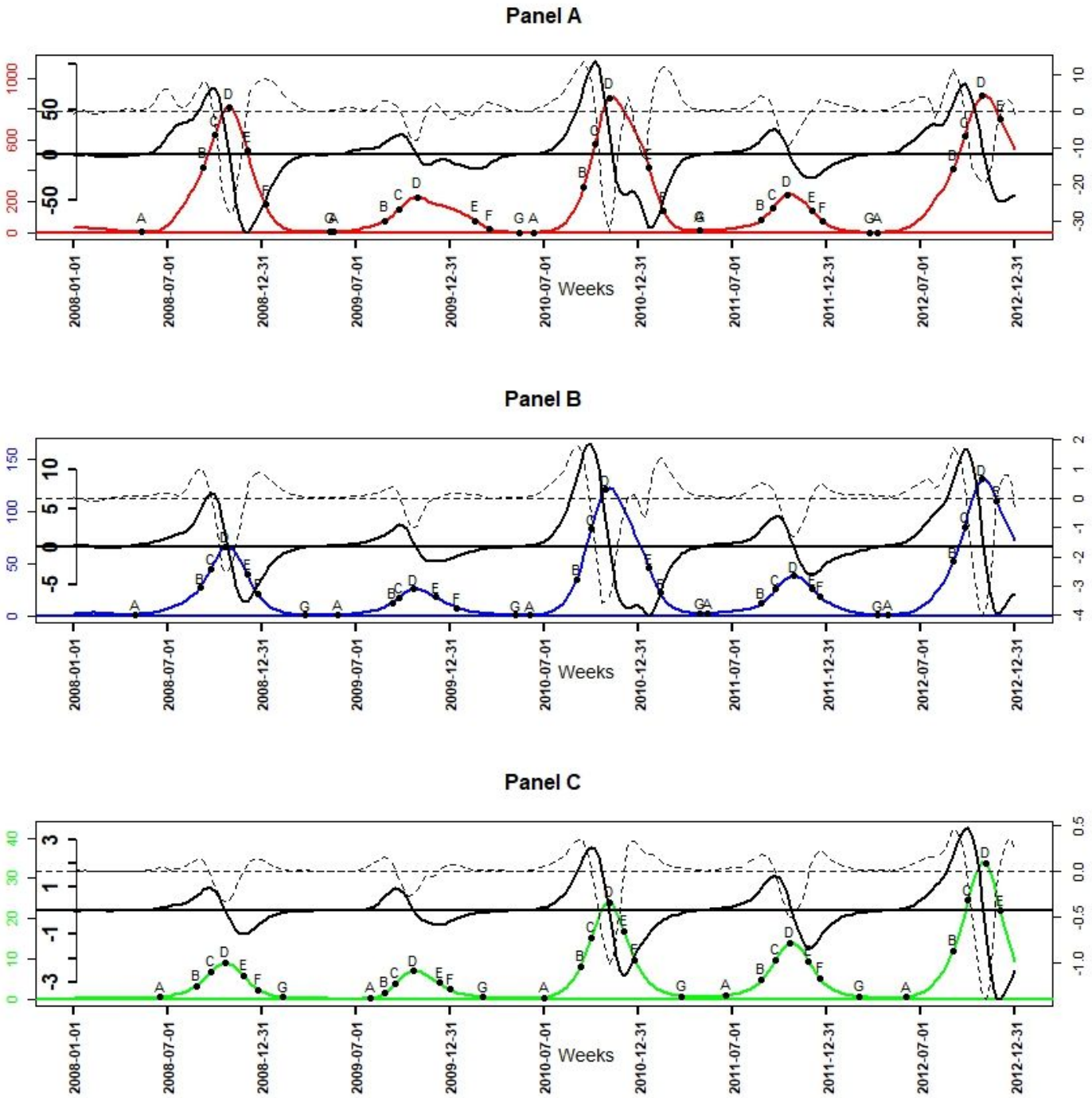

Figure 6

Smoothed incidence in color solid line, their velocity in black bold solid line, their acceleration in black discontinuous line and the epidemiological indicator of their seasonal outbreaks (A: onset, B: near slowdown of growth, C: beginning slowdown of growth, D: Peak, E: beginning acceleration of decline, F: beginning of tail , G: end): high incidence pattern in red (Panel A), intermediate incidence pattern in blue (Panel B) and low incidence pattern in green (Panel C).

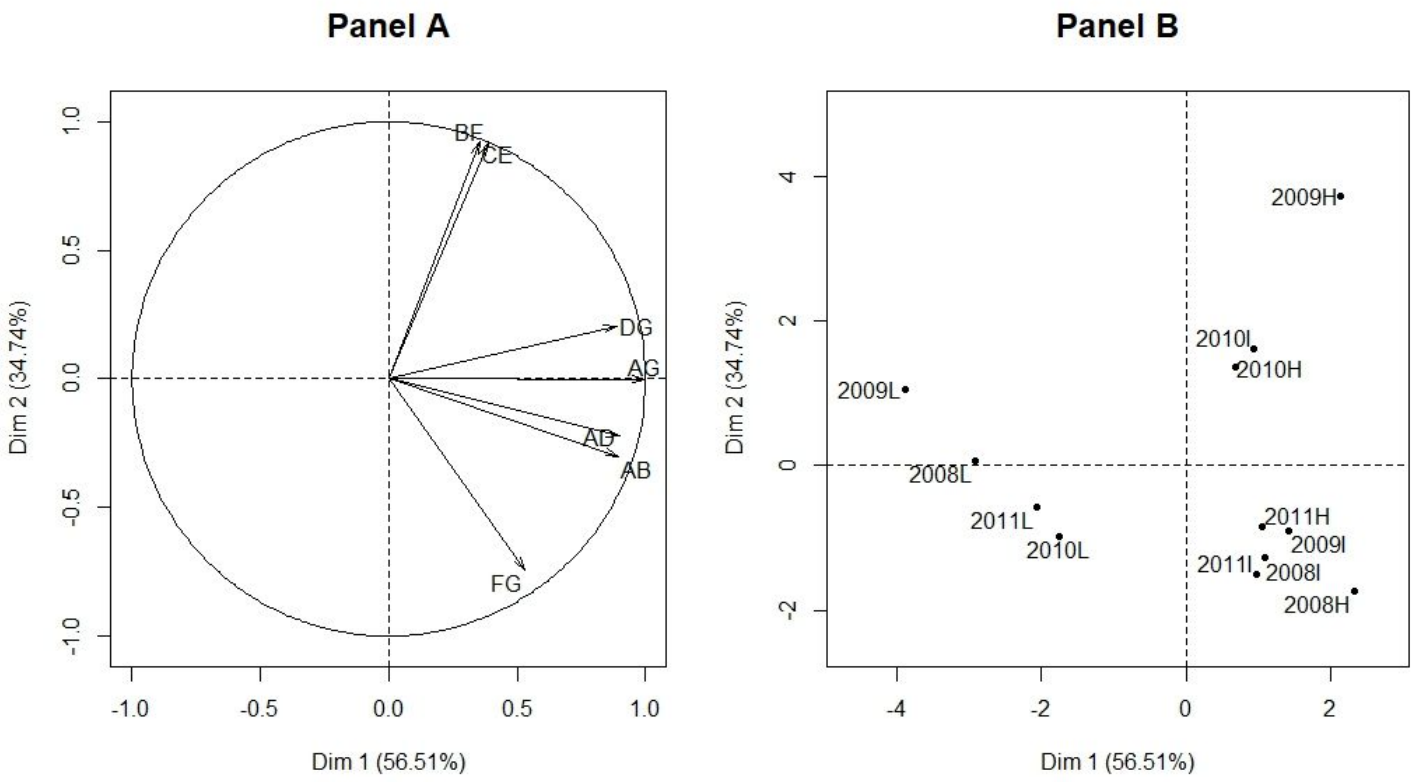

Figure 7

Principal component analysis on duration epidemiological indicator and seasonal outbreaks of the patterns: epidemiological indicator map (Variables, Panel A), seasonal outbreaks of the patterns map (Individuals, Panel B)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile3.jpeg
- Additionalfile2.jpeg
- Additionalfile4.docx