Extracting Most Impacting Emergency Department Patient Flow By Embedding Laboratory-confirmed and Clinical Diagnosis on The Stiefel Manifold

Clément Pealat, Guillaume Bouleux, Vincent Cheutet

Extracting Most Impacting Emergency Department Patient Flow By Embedding Laboratory-confirmed and Clinical Diagnosis on The Stiefel Manifold

Clément Pealat, Guillaume Bouleux, and Vincent Cheutet
Univ Lyon, INSA-Lyon, DISP EA4570,
F-69621, Villeurbanne, France
Email: {clement.pealat, guillaume.bouleux, vincent.cheutet}@insa-lyon.fr

***Abstract*—Emergency departments (ED) in France are jeopardized each winter by the respiratory viruses. To limit the impact of those viruses, it is essential to have a better understanding of their impact on the patient flow. To tackle this, we propose in this work to use in conjunction ICD-10 code and laboratory-confirmed data with the aim of extracting a relevant patient flow. We first take benefice of the almost periodicity of both clinical diagnosis and laboratory-confirmed data and we embed next the underlying time series on the Stiefel manifold. The distance in the Stiefel manifold is finally used to extract clinical codes which are the nearest to the laboratory-confirmed time series. The results reveal that some of the respiratory and cardiac disorders codes have the same behaviours than that of the winter circulating viruses. At least, the Flag mean is employed to dispose of a picture of both the patient flow and the the length of stay for patients who might be infected by winter viruses.**

***Index Terms*—Stiefel Manifold, Flag mean, Subspaces of different dimensions, RSV, Influenza, Patient Flow**

1. Introduction

Emergency departments (ED) are regularly in a situation of overcrowding, especially during winter and viral outbreak [8]. This overcrowding badly impacts the hospital and induces a degradation of the welcome capacity and of the care quality [2].
To avoid this phenomenon, action plans can be set up (increasing the number of beds, staff,..), but they have human and material costs. It is important to detect early and with precision when the overcrowding will occur. Data from the emergency departments can be obtained, and some methods have been applied to predict this overcrowding. For example, Change point detection [5], forecasting [1, 4], and warning alarm have been implemented.
But, those tools can be efficient only if a work on the comprehension of the patient flow in the ED is done beforehand. Indeed, the patient flow is due to a large variety of diagnoses and only a few are linked with winter epidemics. When a patient comes to the ED, he/she is classified by a diagnostic code[1] (ICD), and the virus behind his/her illness can stay hidden. Only, some complementary laboratory tests can detect the virus. To erase the noise due to other diagnoses, clustering algorithms can be applied to determine which diagnoses are relevant. For this type of problems, the distance evaluation in clustering is the key research gap in order to obtain consistent results. In this study, we propose the use of a distance adapted to our problem, which would be used for extracting relevant clinical diagnosis.
The two viruses responsible for the overcrowding during winter are known: there are the Respiratory Syncytial Virus (RSV) and the Influenza [10], and this study applies to both pediatric and adult ED. The flow of pediatric and adult ED are treated in two separate cases because the viruses do not have the same impact on the two populations:

- The principal virus that impacts the pediatric ED is the RSV. For the children, the diagnostic codes are clear: the RSV is associated with the symptoms of the brionchiolitis, and for the influenza, a child come generally directly, and has the ICD code of the influenza (ICD-code J10).

- In the adult ED, the population is impacted mostly by the influenza even if some recent studies [7, 1] have shown that the RSV is also an impacting virus for this population. The associated clinical codes are however not as clear as for the children. Indeed, a patient, with one of these viruses, comes to the ED for some complications linked to the virus, and his/her ICD-code is due to those complications and not the virus by itself.

2. About the Data

In this study, the University Hospital of Saint-Etienne gave us access to ED data. Each week, the number of entries for each ICD, for both pediatric and adult emergency, are registered in the EHR of the Hospital. Moreover, the number of positive tests realized in laboratory for RSV and influenza are included as well. The data are under the form of time series, each one representing the evolution of ICD code, or laboratory-confirmed test over time. In complement, the

PREDAFLU project, French Ministry of health

1. The diagnostic code, defined by WHO, gives the opportunity to classify each diagnostic, with the 10th International Classification of Disease, ICD-10

length of stay is also collected. Those sets of data are from June 2013 to august 2017. Since the data are from a medical field and linked with the ICD (a system used for 18 years), they are well completed, with no missing information, and they are considered as reliable.
The data are composed of an important number of ICD diagnoses. It has been decided not to keep diagnostics with a weak occurrence (less than the average occurrence of a diagnosis during the observation period), and 145 ICD-codes are kept. Finally, to prepare for the study, the data have been standardized.

3. Distance on the Stiefel Manifold

3.1. Embedding

First, a trajectory matrix is defined for each time series [6]. Considering a time series $y_i = (y_i(0), ..., y_i(p-1))$, we subdivide y_i into vectors of size m defined by $Y_i(k) = (y_i(k), ..., y_i(k+m-1))$, $k = 0, \dots, p-m+1$. By concatenating the vectors $Y_i(k)$ for all k, the trajectory matrix Y_i of the time series y_i is finally defined as:

$$Y_i = \begin{bmatrix} y_i(0) & y_i(1) & \cdots & y_i(p-m) \\ y_i(1) & y_i(2) & & \vdots \\ \vdots & \vdots & & \vdots \\ y_i(m-1) & y_i(m) & \cdots & y_i(p-1) \end{bmatrix}$$

This defines a Hankel matrix, and for a better representation of the time series, the classic value used for m is $m = \lfloor p/2 \rfloor$ where p is the length of the time series [6]. Here, the time series present a seasonal behavior of one year, and are 4-years length. Each time series y_i is divided in four time series of one-year length, denoted $Y_i^{(k)}$, $k = 1, \dots, 4$. We obtain therefore the matrix $X_i = \begin{bmatrix} Y_i^{(1)} & Y_i^{(2)} & Y_i^{(3)} & Y_i^{(4)} \end{bmatrix}$ which represents the dynamics of a time series y_i in the embedded space. To gain in stability, and to take into account of the periodical second order statistics of the trajectories, the covariance matrix $R_{X_i} = \frac{1}{4(p-m+1)} \sum_{i=1}^{4(p-m+1)} (X_i - \bar{X}_i)(X_i - \bar{X}_i)^T$ of X_i is used to geometrically represent our data. R_{X_i} is an element of $\mathbb{R}^{m \times m}$ and account now of the seasonal statistical characteristics.

3.2. Determination of the lower dimensional manifold

Each time series is now represented by the matrix R_{X_i}. We now reduce the number of dimensions to cut out noise and redundancy of the data. This reduction allows us to characterize the lower-dimensional manifold which best represents the data. For this reduction, we use the kernel principal component analysis (KPCA), with the Gaussian Radial Basis Function :

$$K(R_{X_i}, R_{X_j}) = exp\left(-\frac{\|R_{X_i} - R_{X_j}\|^2}{\sigma^2}\right)$$

We have decided to keep only the first n components such as:

$$\frac{\sum_{i=1}^{n} \lambda_i}{\sum_{i=1}^{m} \lambda_i} > 0.9$$

with lambda eigenvalues of the centered kernel matrix in decreasing order. Thus, each time series y_i is now defined by the matrix $D_i \in \mathbb{R}^{m \times n}$, reduction of R_{X_i}.

3.3. Stiefel manifold and principal angles

At this step, it worth noting to recall that the original time series are now embedded in a m-dimensional space whose representation is a collection of n orthogonal vectors in that space. This precisely leads to the Stiefel manifold, a submanifold of $\mathbb{R}^{m \times n}$ defined by $V_n(\mathbb{R}^m) = \{A \in \mathbb{R}^{m \times n} : A^T A = Id\}$, which is also seen as an homogeneous space defined as $V_n(\mathbb{R}^m) \simeq O(m)/O(m-n)$ with $O(m)$ the orthogonal group in dimension m. Two metrics can be used, they are defined on the tangent space of $V_n(\mathbb{R}^m)$ with either the euclidean metric

$$< A, B >= Tr(A^T B) \tag{1}$$

coming from the definition of $V_n(\mathbb{R}^m)$ as a submanifold of $\mathbb{R}^{m \times n}$ and the canonical metric for $Q \in V_n(\mathbb{R}^m)$

$$< A, B >= Tr\left(A^T \left(I - \frac{1}{2} Q Q^T\right) B\right) \tag{2}$$

coming from the homogeneous space definition of $V_n(\mathbb{R}^m)$. The benefice of using the Stiefel manifold embedding is the correspondence of these two metrics regarding the geodesic distance between the points of $V_n(\mathbb{R}^m)$. Indeed, both yields to the same geodesic distance given by the principal angles of their subspaces in such a way that for any $A, B \in V_n(\mathbb{R}^m)$, we have

$$distance(A, B) = \left(\sum_{i}^{n} \theta_i^2\right)^{\frac{1}{2}} \tag{3}$$

with θ_i, $i = 1, \dots, n$ the principal angles between A and B. However, in our case the D_i matrices have different dimensions and they belong to different Stiefel manifolds. It is then necessary to modify the usual metrics defined by (1) and (2) for accounting this. This is proposed for example in [11], where the authors have shown that is enough to complete the lower dimensional space by orthogonal vectors in order to obtain a metric. The distance (3) is then adjusted such as for $A \in V_k(\mathbb{R}^m)$, $B \in V_l(\mathbb{R}^m)$ and $k < l$

$$distance(A, B) = \left((l-k)\pi^2/4 + \sum_{i}^{k} \theta_i^2\right)^{\frac{1}{2}} \tag{4}$$

4. Results

In table 1, are reported the distances for the viruses and the three closest clinical codes, for the children and the adults respectively. The ICD codes J correspond to

Population	Virus	Code	Distance
Adult	RSV	J20	3.188949
		J18	3.857907
		I10	4.161904
	Influenza	J20	3.166252
		J18	3.852630
		I10	4.110217
Child	Influenza	J10	0.200096
		J11	0.361659
		J04	0.790983
	RSV	J21	0.055627
		J00	0.327844
		J04	0.723412

TABLE 1. TABLE OF THE CLOSEST DISTANCES TO BOTH RSV AND INFLUENZA VIRUS FOR THE ADULT AND PEDIATRIC ED BASED ON THE DISTANCE OF (4).

respiratory symptoms, with J10, J11 the Influenza and J21 Bronchiolitis.

As expected for the children, the clinical codes are clear. J11 and J10 are the estimated nearest clinical codes from the laboratory-confirmed Influenza, J21 is the nearest to the RSV. For example, the distance between RSV and J04 is more than ten times the one between RSV and J21. For the adults, J20, J18 and I10 are the closest codes for both of the viruses. The precedent study [9], that has worked on the same data, gives similar results, especially for the children. For the adults, some other diagnostics have been detected, especially more cardiac ones (ICD codes I). This an ongoing work to used the metric proposed in this work for adapting clustering methods.

4.1. Reconstruction of the time series

Depending on the gap between the distances, we have kept up to three clinical codes. The patient flow is then reconstructed regarding these selected time series. The principle is to average the associated D_i matrices and next orthogonaly project all the clinical codes of the EHR on the subspace spanned by the obtained and averaged matrix. Due to the lack of space, we can not go further on the underlying problem of averaging orthogonal bases of different dimensions. Again, some hints could be found in the reference [3] which uses the flag mean.

In figure 1-(a), we have plotted the total number of arrivals (all ICD code combined) in the pediatric emergency department and the number of laboratory-confirmed RSV for the children. From this, it is not possible to apply method of detection for the winter epidemic, the data are too corrupted and nothing clearly emerged. In figure 1-(b), the reconstructed patient flow associated with RSV seems to perfectly match the laboratory-confirmed time series. The chronic winter peaks of the reconstructed patient flow and that of the laboratory-confirmed RSV appear simultaneously and with the same proportions. We proposed the same study for the adult population. Even if the patient has been refined regarding that of the Figure 2-(b), there is still a mislead between RSV and Influenza for the adults. Nevertheless, the reconstructed patient flow of Figure 2-(b) can significantly help in starting winter plan.

Our last analysis, focuses on the overcrowding of the ED. In figure3, is displayed the percentage of the length of stay associated with the three clinical codes associated with Influenza reported in Table 1 among the total ED length of stay. Clearly, we can see that during the peak of the epidemics, up to 70% of the length of stay is explained by those three clinical codes. This reinforced the good clinical codes extraction performed by the distance on the Stiefel mannifold and the Flag average of the related subspaces.

5. Conclusion

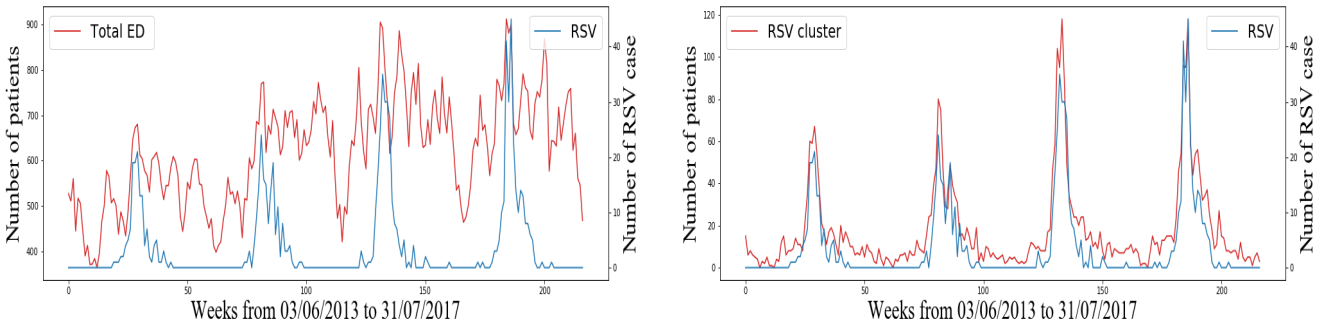
Emergency departments (ED) in France are largely impacted each winter by the respiratory viruses. To limit the impact of those viruses on the crowding and on the mortality at the ED, it is crucial to have a way to catch a picture of the circulation for these virus on the patient flow. This is precisely what is proposed in this work. We have used the laboratory-confirmed database for RSV and Influenza virus of the Saint Etienne Hospital, France in conjunction with all the arrivals recording during 4 years. The underlying time series have been embedded into the Stiefel manifold in order to compute a suitable distance and extract then the three nearest ICD-codes from RSV and Influenza. Since times series belong to different Stiefel manifolds, we have computed a new distance based on Schubert varieties. The results reveal that some of the respiratory and cardiac disorders codes have the same behaviours than that of the winter circulating viruses, especially for children. In order to have an idea on the associated patient flow, we have also computed a Flag mean to reconstruct the associated patient flow. With this, this new times series becomes an input for detection procedure and might reduce the crowding at the ED. Finally, we have observed the length of stay for the reconstructed times series of Influenza and showed that up to 70% of the total length of stay at the ED was explained by the associated clinical diagnoses.

References

[1] G. Bouleux, E. Marcon, and O. Mory. “Early index for detection of pediatric emergency department crowding”. In: *IEEE journal of biomedical and health informatics* 19.6 (2015), pp. 1929–1936.

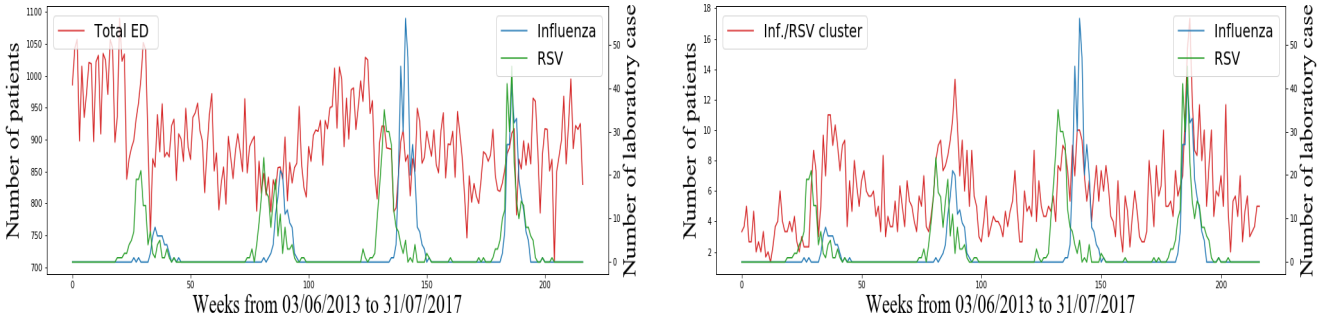
[2] R. W. Derlet and J. R. Richards. “Overcrowding in the nations emergency departments: complex causes and disturbing effects”. In: *Annals of emergency medicine* 35.1 (2000), pp. 63–68.

[3] Bruce Draper et al. “A flag representation for finite collections of subspaces of mixed dimensions”. In: *Linear Algebra and its Applications* 451 (2014), pp. 15 –32. ISSN: 0024-3795.



(a) (b)

Figure 1. (a)-Number of arrivals at the pediatric ED and laboratory-confirmed RSV, (b)-reconstructed patient flow for clinical codes which are the nearest to the RSV confronted with the laboratory-confirmed RSV.



(a) (b)

Figure 2. (a)-Number of arrivals at the adult ED and laboratory-confirmed RSV plus Influenza, (b)-reconstructed patient flow for clinical codes which are the nearest to the Influenza confronted with the laboratory-confirmed RSV plus Influenza.

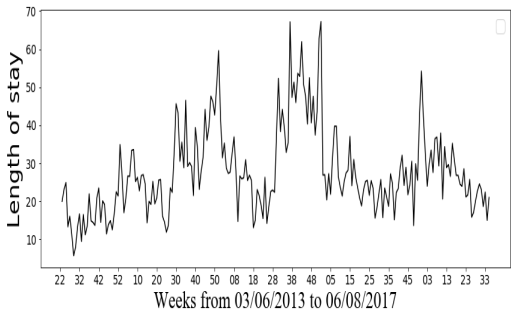


Figure 3. Percentage of length of stay due to the cluster

[4] M. Dugast et al. "Improving Health Care Management Through Persistent Homology of Time-Varying Variability of Emergency Department Patient Flow". In: *IEEE journal of biomedical and health informatics* (2018).

[5] Taha A Kass-Hout et al. "Application of change point analysis to daily influenza-like illness emergency department visits". In: *Journal of the American Medical Informatics Association* 19.6 (2012), pp. 1075–1081.

[6] C. O'Reilly, K. Moessner, and M. Nati. "Univariate and Multivariate Time Series Manifold Learning". In: *Knowledge-Based Systems* 133 (2017), pp. 1–16.

[7] D. L. Schanzer, J. M. Langley, and T. WS Tam. "Role of influenza and other respiratory viruses in admissions of adults to Canadian hospitals". In: *Influenza and Other Respiratory Viruses* 2.1 (2008), pp. 1–8.

[8] D. L. Schanzer and B. Schwartz. "Impact of seasonal and pandemic influenza on emergency department visits, 2003–2010, Ontario, Canada". In: *Academic Emergency Medicine* 20.4 (2013), pp. 388–397.

[9] G. Soler et al. "Emergency Department Admissions Overflow Modeling by a Clustering of Time Evolving Clinical Diagnoses". In: *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*. IEEE. 2018, pp. 365–370.

[10] F. HY Yap et al. "Excess hospital admissions for pneumonia, chronic obstructive pulmonary disease, and heart failure during influenza seasons in Hong Kong". In: *Journal of medical virology* 73.4 (2004), pp. 617–623.

[11] K. Ye and L-H Lim. "Schubert varieties and distances between subspaces of different dimensions". In: *SIAM Journal on Matrix Analysis and Applications* 37.3 (2016), pp. 1176–1197.