A TOPOLOGICAL DATA ANALYSIS APPROACH TO THE EPIDEMIOLOGY OF INFLUENZA

Joao Pita Costa and Primož Škraba
Artificial Intelligence Laboratory, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773144; fax: +386 1 4773905
e-mail: joao.pitacosta@ijs.si

ABSTRACT

Influenzanet is a system to monitor the activity of influenza-like-illness [ILI] with the aid of internet volunteers. Topological data analysis [TDA] examines the structure of data and contributes to the development of medicine, studying properties of a continuous space by the analysis of a discrete sample of it. Using TDA we analyze the topology of Influenzanet data identifying noise and distinguishing higher dimension features. This is done both in terms of the overall structure of a disease as well as its evolution. It provides a way to test agreement at a global scale arising from standard local models. We also compare this qualitative method to other quantitative methods such as Fourier analysis or dynamical time warping [DTW].

1 INTRODUCTION

Topological data analysis [TDA] provides us with the topological features that describe the structure of a given point cloud. It infers high-dimensional structure from low-dimensional representations and studies properties of a continuous space by the analysis of a discrete sample of it, assembling discrete points into global structure. The basic technique encodes topological features of a given point cloud by diagrams representing the lifetime of those topological features. A good introduction to topological data analysis can be found in [1]. Recently, these topological methods on data have seen a relevant application to the study of the influenza virus as described in [2].

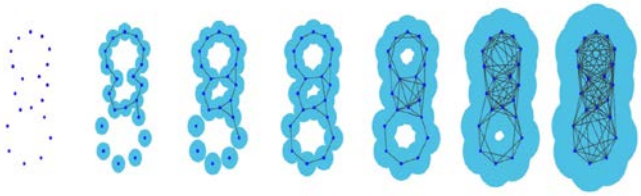

Figure 1. Topological data analysis: the filtration of a simplicial complex of a given pointcloud according to the growing radious of balls centered in the input data points.

The system *Influenzanet* monitors online the activity of *influenza-like-illness* [ILI] with the aid of volunteers via the internet. It has been operational for more than 10 years, and at the EU level since 2008. Influenzanet obtains its data directly from the population, contrasting with the traditional system of sentinel networks of mainly primary care physicians. Influenzanet is a fast and flexible monitoring system whose uniformity allows for direct comparison of ILI rates between countries [5].

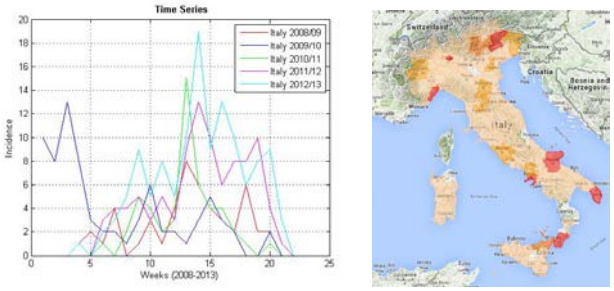

Figure 2. Influenzanet: the time-series for the incidence of influenza in Italy during the flu seasons of 2008-2013 (on the left); a screenshot of the influenzanet system in Italy, taken in May 2015 (on the right).

Our goal with this project is to analyze the Influenzanet data using persistence, identifying topological features relevant to the epidemiological study. To do so, we identify data noise, distinguish higher dimension features and look at the overall structure of the disease as well as its evolution during the flu season in Portugal and Italy. In particular, this provides a way to test agreement at a global scale arising from standard local models.

2 TOPOLOGICAL ANALYSIS OF EPIDEMIOLOGICAL DATA.

The *Mahalanobis distance* is a measure of the distance between a point *P* and a distribution *D*, widely used in cluster analysis and classification techniques. When considering this metric on the space while using TDA, we get a perspective of that space under different scales, where small features will eventually disappear. We have used in [8] several techniques to preprocess the input data, including subsampling and colliding data points that are closer than a

given parameter. In particular, we embed the data in higher dimensions, compute persistence, and look for outliers.

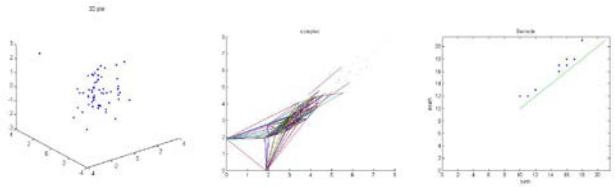

Figure 3. The pipeline for the computation of topological data analysis for the time series of Italy 2009/10: the given pointcloud of the input data (on the left); the Viatoris-Rips complex approximating the space of the pointcloud (in the center); the correspondent persistence diagram encoding the lifetime of the persistent topological features (on the right).

The analyzed data lists the number of active participants and the number of ILI onsets, for three different ILI case definitions of the Influenzanet in Italy for every week in years of the Influenza seasons from 2010/11 to 2012/13. Based on this data we have used several algorithms to preprocess it, prior the construction of the Vietoris-Rips complex that corresponds to the given data. This method permits us to encode the qualitative features of that data into a persistence diagram.

The images in Figure 3 show the cloud of input data points, the corresponding simplicial complex, and persistence diagram for dimension 1. These topological tools complement the information obtained by classical data analysis. The computation of the persistence diagrams is done via Vietoris-Rips complexes using *Perseus*, the open source persistent homology software [4]. The input structure is given as a symmetric distance matrix where the entries come from pairwise distances between points in a given point cloud. In the figures below we can see three steps of the construction of the Vietoris-Rips complex that will provide us with the persistence diagram encoding the topological information of the Influenzanet data.

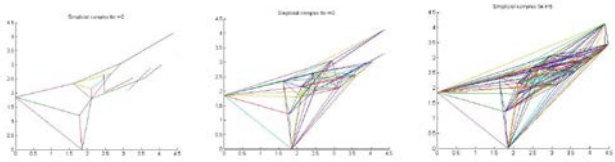

Figure 5. The filtration of the simplicial complex at several levels varying according a parameter *r* for the input time series of Italy in the flu season of 2009/2010: $r = 2$ (on the left); $r = 3$ (in the center); $r = 5$ (on the right).

3. QUANTITATIVE AND QUALITATIVE ANALYSIS OF INFLUENZA.

Fourier analysis is widely used to identify patterns in a time series. We used the time series of the incidence of influenza in Portugal and Italy for the flu seasons of 2008-2013. In Figure 4 we can see the plot of the two time series and their correspondent Fourier transform.

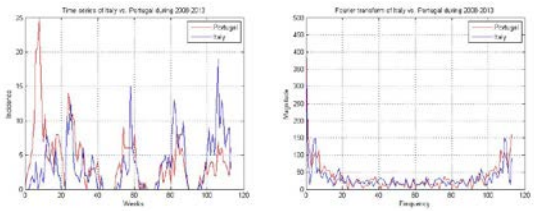

Figure 4. Comparing the flu seasons of Portugal and Italy during 2008-2013: the time series (on the left); the Fourier transform (on the right).

We computed in [9] the Fourier transform for each pair of time series *(country, year)* to compare the flu seasons of Portugal and Italy. In that work we compared the quantitative methods of Fourier analysis with the qualitative methods of TDA. In Figure 7 the reader can see an extension of the results of this comparison with highlighted biggest and smallest values.

When comparing two time series that may vary in time or speed it is usual to apply the algorithm *dynamic time warping* [DTW] measuring the similarity between those temporal sequences. In this study we compared each pair of time series *(country, year)* obtaining the respective measure that can be seen in the table of Figure 7.

The usage of TDA for the analysis of time series was explored in [6] towards the quantification of periodicity and identification of periodic signals in gene expression in [7]. We also use TDA to analyze the input time series data, following an approach developed specifically for Influenza. Barcodes and correspondent persistence diagrams seen as multi-scale signatures encode the lifetime of topological features within pairs of numbers representing birth and death times. We have computed a persistence diagram for each time series *(country, year)* embedded in higher dimensions. As shown by the persistence diagrams below, the distinguishable features are seen in dimension 1.

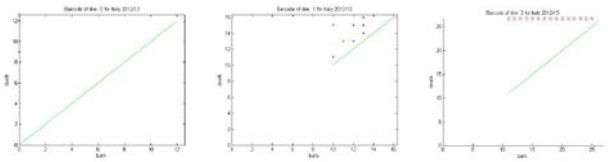

Figure 6. The persistence diagrams for the input time series of Italy in the flu season of 2009/2010: dimension 0 (on the left); dimension 1 (in the center); dimension 2 (on the right). The red circles mean that the lifetime of the considered features does not end.

Persistence landscapes are techniques of TDA that permit us to measure the pairwise distance between persistence diagrams at several different levels. The distance value between these two persistence diagrams in the tables of Figure 7 was calculated using the *persistence landscapes*

toolbox [3] to compute the distance between diagrams considering different norms.

The following tables represent the comparison between the Fourier analysis, dynamical time warping and topological analysis of the incidence of influenza in Italy and Portugal for the flu seasons of 2008-2013.

Fourier		Portugal				
		2008	2009	2010	2011	2012
Italy	2008	**2,2518**	1,523	2,0147	1,1977	0,95957
	2009	1,523	1,0536	1,3576	0,86667	0,74338
	2010	2,0147	1,3576	1,1635	0,70203	0,67165
	2011	1,1977	0,86667	0,70203	0,67071	0,6352
	2012	0,95957	0,74338	0,67165	0,6352	**0,61559**

DTW		Portugal				
		2008	2009	2010	2011	2012
Italy	2008	89	80	20	15	15
	2009	106	32	58	60	58
	2010	75	88	**13**	23	23
	2011	60	92	16	28	35
	2012	44	**111**	35	48	55

TDA		Portugal				
		2008	2009	2010	2011	2012
Italy	2008	2,91548	2,85774	2,25462	2,81366	2,51661
	2009	2,32737	2,27303	1,73205	1,63299	1,75594
	2010	**0,288675**	0,408248	1,22474	1,32288	0,957427
	2011	0,957427	1	1,35401	1,52753	1,32288
	2012	4,09268	4,04145	3,37886	3,7305	**3,58236**

Figure 7. Comparing the flu seasons of Portugal and Italy during 2008-2013: the distance tables for the Fourier analysis (on the top), the dynamic time warping (on the center), and the topological data analysis (on the bottom).

When comparing the distances obtained by Fourier analysis, DTW and TDA we can see that these three methods look at different features of the data.

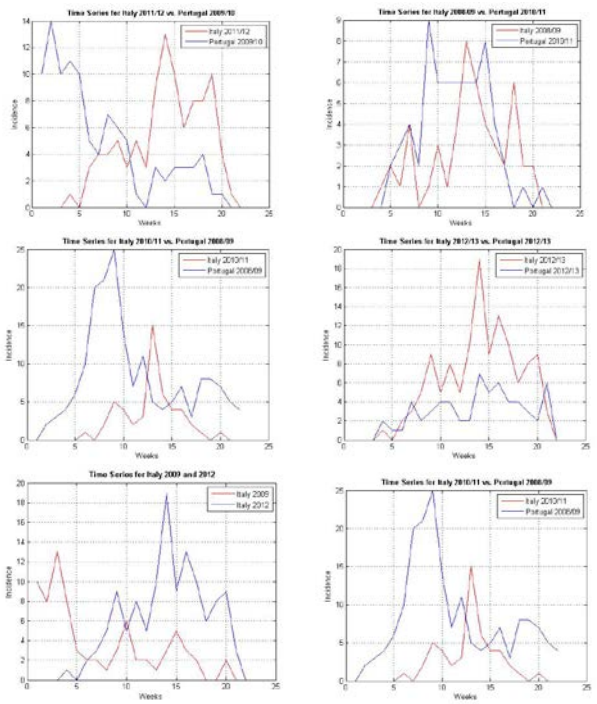

Figure 8. Comparing the flu seasons of Portugal and Italy during 2008-2013: selected plots of time-series to compare the results in the Fourier analysis, the topological data analysis and the dynamical time warping.

The plots in Figure 8 represent time-series for selected flu seasons from 2008 to 2013. They serve us to compare the different data analysis methods used in this study.

When comparing the distances between Italy 2011/12 and Portugal 2009/08, the Fourier provides us with a high value of 92, while DTW analysis has a low value of 0,86667. On the other hand, for the flu seasons of Italy 2008/09 and Portugal 2010/11, the DTW has a low value of 20 while the Fourier analysis has a relatively high value of 2,0147. In the first case the monotony of the curves match, although the periodicity not being close. The second case shows two high peaks for Italy 2008/09 against one for Portugal 2010/11 explaining the low level of DTW.

To compare the quantitative Fourier analysis with the qualitative analysis of TDA we look at the flu seasons of Italy 2010/11 and Portugal 2008/09 where TDA achieved the low value of 0.288675 and the Fourier analysis reached the high value of 2,0147. On the other hand, the flu seasons of Italy and Portugal in 2012/13 reach a high TDA value of 3,58236 and a low Fourier value of 0,61559. The first case shows a big difference of peaks which does not happen in the second case where the periodicity is lower, implying the lower level for the Fourier analysis.

Finally, the comparison between TDA and DTW points us to the flu seasons Italy 2008/09 and Portugal 2012/13, where TDA reached a high value of 2,51661 (due to the higher similarity of peaks) and DTW reached a low value of 15 (describing the different behavior of the curves); and the flu seasons of Italy 2010/11 and Portugal 2008/09, where TDA achieved a low level of 0.288675 (with great difference of peaks as pointed out earlier) and DTW achieved the high value 75 (pointing out the similar behavior of the curves).

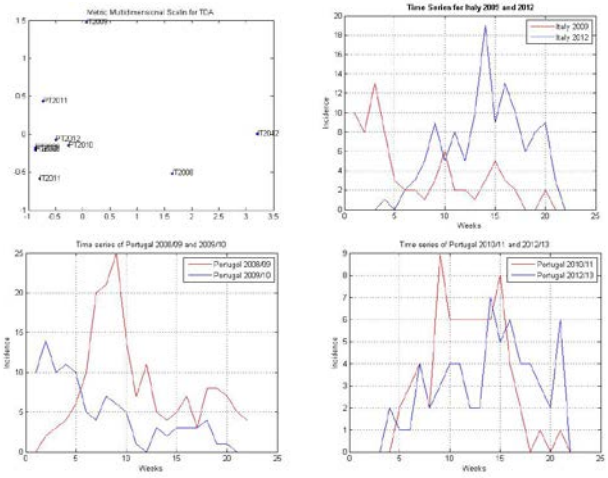

Figure 9. Comparing the flu seasons of Italy and Portugal during 2008-2013 using metric multidimensional scalling (on the upper left) to identify:

the outlier flu seasons of Italy 2009/10 and 2012/13, with time series ploted for analysis and interpretation (on the upper right); the close flu seasons of Portugal 2008/09 and 2009/10 (on the lower left); and the flu seasons of Portugal 2010/11 and 2012/13, close to the diagonal (on the lower right).

We used multidimensional scaling as in Figure 9 to identify outliers for each of the three methods within the flu seasons analyzed in this study. TDA provides a qualitative analysis of the time series of the incidence of influenza, looking in particular at the peaks and dramatic changes. In that perspective, the time series of Italy 2009/10 and 2012/13 plotted in Figure 9 describe very different flu seasons with very different peaks. On the other hand, the flu seasons of Portugal 2008/09 and 2009/10 are identified being very close with very similar peaks, although the behavior of the curve being different. The knowledge on secondary attack rates in the influenza season is of importance to access the severity of the seasonal epidemics of the virus, estimated recently with information extracted from social media in [10]. Here lies a strong point of TDA where it can provide relevant contribution complementing other methods.

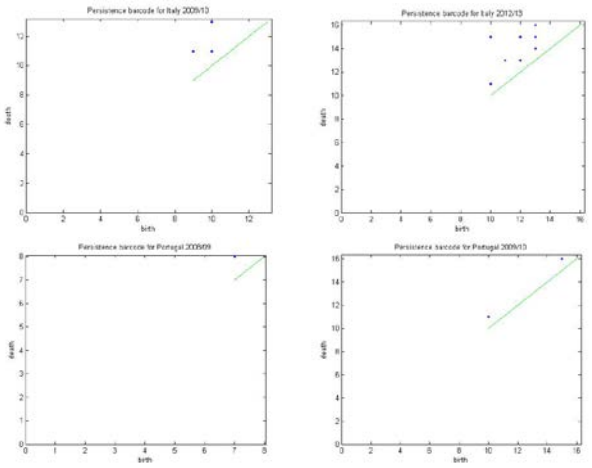

Figure 10. Comparing the flu seasons using persistence diagrams for dimension 1 for: Italy 2009/10 (on the upper left), Italy 2012/13 (on the upper right), Portugal 2008/09 (on the lower left), and Portugal 2009/10 (on the lower right), identified as particular cases in Figure 9.

The persistence diagrams of Figure 10, correspondent to the identified flu seasons of Italy 2009/10 and Italy 2012/13, and Portugal 2008/09 and 2009/10. They encode the lifetimes of the topological features of the curves of the time series of those seasons. Persistence diagrams are a clear and practical tool that allows us the detection of outliers and to capture the qualitative features of the dynamics of the system. These ideas provide a new approach to the analysis of the seasons in the epidemiology of Influenza.

4. CONCLUSION AND FUTURE WORK

The study of Epidemiology is a great source of problems relating to nonlinear systems, large scale data and development of more accurate models, where TDA can contribute, providing high dimension techniques for medical data analysis. In this study we showed how they can be used to analyze the incidence for different ILI case definitions, contributing to a better understanding of the features distinguished by those definitions. The information provided by quantitative methods such as DTW or the Fourier analysis of time series can be complemented by the topological analysis of that data. The examples considered in Figure 8 show that these methods do not express the same information about the development of the epidemics during the flu season. The knowledge provided by each of these methods complements the knowledge coming from the other methods and can be put together in a global information map. Further research considers the analysis of the impact of the qualitative aspects of TDA for modeling and prediction of the current Influenza season. We will also use state of the art artificial intelligence methods to learn metrics more appropriate to the input time series data aiming, to grasp a better understanding of the severity of the epidemics both in past seasons and during the ongoing season.

ACKNOWLEDGMENTS

The authors would like to acknowledge that his work was funded by the EU project TOPOSYS (FP7-ICT-318493).

REFERENCES

[1] G. Carlsson (2009). Topology and data. Bulletin of the American Mathematical Society 46.2 : 255-308.

[2] J. M. Chan, G. Carlsson and R. Rabadan (2013). Topology of viral evolution. Proceedings of the National Academy of Sciences 110.46: 18566-18571.

[3] P. Dlotko (2014). Persistence Landscapes Toolbox (www.math.upenn.edu/ dlotko).

[4] V. Nanda (2014). Perseus (www.sas.upenn.edu/˜ vnanda/perseus).

[5] D. Paolotti et al (2014). Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clinical Microbiology and Infection 20.1: 17-21.

[6] J. A. Perea and J. Harer (2013). Sliding windows and persistence: An application of topological methods to signal analysis. Foundations of Computational Mathematics 15.3: 799-838.

[7] J. A. Perea et al (2015). SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. BMC bioinformatics 16.1: 257.

[8] J. Pita Costa and P. Škraba (2014). A topological data analysis approach to epidemiology. European Conference of Complexity Science 2014.

[9] J. Pita Costa and P. Škraba (2015). Topological epidemiological data analysis. ACML Health 2015.

[10] E. Yom-Tov et al (2015). Estimating the Secondary Attack Rate and Serial Interval of Influenza-like Illnesses using Social Media. Influenza and other respiratory viruses. DOI:10.1111/irv.12321