A TOPOLOGICAL DATA ANALYSIS APPROACH TO INFLUENZA-LIKE-ILLNESS

Joao Pita Costa[1 2 3], Primož Škraba[3 4], Daniela Paolotti[5] and Ricardo Mexia[6]

***Abstract*—Influenzanet is an online automated system to monitor the activity of influenza-like-illnesses (ILI) with the aid volunteers through the internet. The discussion in this paper has focus on the topological analysis of the Influenzanet dataset, examining the structure of that data to provide insights on the behaviour of the ILI season and comparing ILI seasons. The general approach performs a qualitative analysis based on the topology of the curves of the time series generated by each ILI season. It provides a way to test agreement at a global scale arising from local models. We also show the complementary potential of this qualitative method to quantitative methods such as Fourier analysis and dynamical time warping.**

***Keywords*—digital epidemiology, influenza-like-illness, influenzanet, persistent homology, computational topology, persistence diagram.**

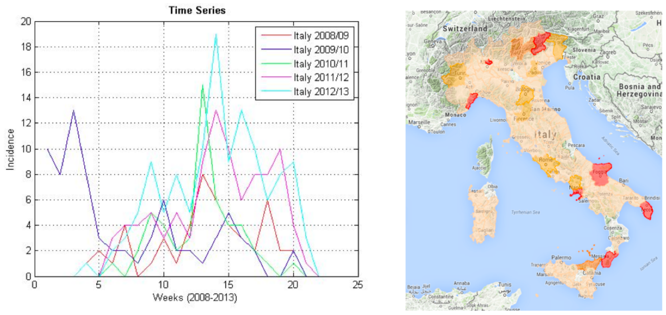


Fig. 1. The curves of the time series of ILI incidence in Italy during the ILI seasons of 2008/09-2012/13 (on the left); a screenshot of the influenzanet system in Italy, in May 2015 (on the right), where the infection level based on reported symptoms goes from ocher (min) to red (max).

I. Introduction

Due to the pandemic potential of influenza, a complete knowledge of the development of each ILI season is a public health priority. In this paper we contribute to that aim by discussing the problem of comparing influenza seasons throughout the years (selected countries: Portugal and Italy) based on the topological behavior of the curve of the time series of incidence in the population. We also discuss the identification of recurrence that would correspond to patterns in the influenza season. To do that we recur to the novel methods of topological data analysis [TDA], providing us with the persistent topological features that describe the structure of the data. The basic technique encodes topological features of a given point cloud by diagrams representing the lifetime of those topological features (see Figure 1). A good introduction to topological data analysis can be found in [2]. Topological methods on data have seen application to the study of the influenza viral evolution in [3] and other public health priorities such as diabetes [5] or cancer [7]. Our goal in this paper is to analyze the Influenzanet data using persistent homology (i.e. *persistence*), identifying persistent topological features relevant to the digital epidemiology study. The *Influenzanet* system monitors the activity of *influenza-like-illness* [ILI] in Europe with the aid of online volunteers. It has been operational in Portugal since 2005, and in Italy since 2008. Influenzanet obtains its data directly from the population, contrasting with the traditional system of sentinel networks of mainly primary care physicians [8]. Influenzanet was shown to be a fast and flexible monitoring system whose uniformity allows for direct comparison of ILI rates between countries [13]. In this paper we look at the overall structure of several influenza seasons as well as their evolution in Portugal and Italy. In particular, this provides a way to test agreement at a global scale arising from standard local models. The method for comparing time series data through TDA, is innovative in the context of the analysis of ILI seasons. It differs from other approaches by providing us with a tool that is independent of the different sizes of the samples collected in each country, comparing the shape of the data generated in each ILI season between countries. We will compare it with dynamic time warping [DTW], that can also compare the time series based on their behaviour, independent from the variations in time. A complementary study is to look for periodicity in the ILI season. The usage of TDA for the analysis of time series was explored in [9] towards the quantification of periodicity and identification of periodic signals in gene expression. The method 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. Similarly, using TDA to analyze the input time series data, after a delay embedding of the time series in R^2 as in [9], we can study periodicity in the Influenzanet data [10], or compare the persistent features of the curves generated by that data [12]. We compare the results with those of Fourier analysis, the standard quantitative analysis of periodicity in the data.

II. TOPOLOGICAL ANALYSIS OF EPIDEMIOLOGICAL DATA

Given the time series of ILI incidence defined by pairs (country, year), we aim to compare them through the analysis of the persistence of topological features. In particular, we

[1]University of Rijeka, Croatia; [2]Quintelligence, Slovenia; [3]Institute Jožef Stefan, Slovenia; [4]University of Primorska, Slovenia; [5]ISI Foundation, Italy; and [6]Instituto Nacional de Saúde Dr. Ricardo Jorge, Portugal

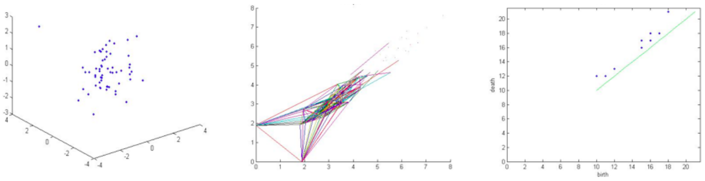


Fig. 2. The pipeline for the computation of topological data analysis for the time series of Italy for 2009/10: the given pointcloud of the input data (on the left); the Viatoris-Rips complex approximating the space of the pointcloud (in the center); and the correspondent persistence diagram encoding the lifetime of the topological features (on the right).

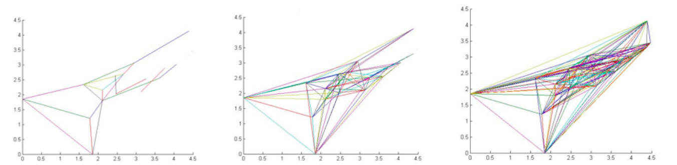


Fig. 3. The filtration of the simplicial complex at several levels varying according a parameter r for the input time series of Italy in the ILI season of 2009/2010: $r = 2$ (on the left); $r = 3$ (in the center); $r = 5$ (on the right).

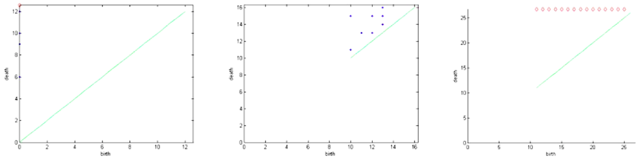
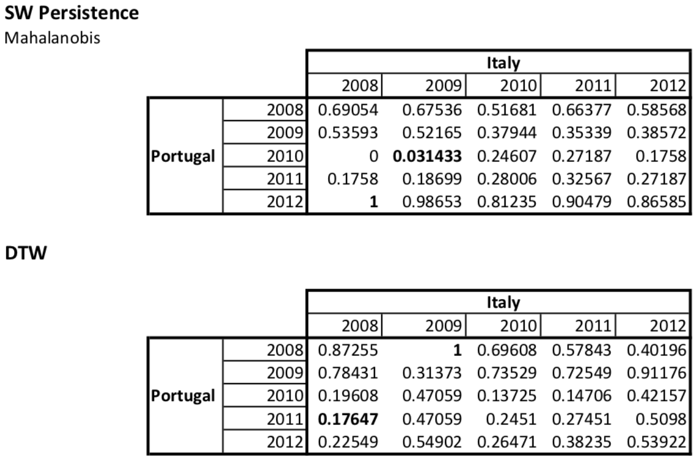


Fig. 4. The persistence diagrams for the input time series of Italy in the ILI season of 2009/10: dimension 0 (on the left); dimension 1 (in the center); dimension 2 (on the right). The red circles mean that the lifetime of the considered features does not end.

embed the data in higher dimensions, compute persistence, and distinguish data noise and outliers. In that, we get a perspective of that space under different scales, where small features will eventually disappear. The approach in [14] using the Takens Delay Embedding Theorem permits us to project the time series data onto a n-dimensonal space and from it construct a persistence diagram corresponding to the time series of influenza incidence. 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. As persistence is independent from the metric used, in this paper we consider the Mahalanobis metrics on the space to construct a *simplicial complex* (i.e. a combinatorial approximation of the space based on points, line segments, triangles, and their n-dimensional counterparts [2]) within the TDA analysis pipeline, represented in Figure 2. The complexity of computations grows fast with the rise of input data due to the usage of these topological structures. For the sake of efficiency we have used several methods to preprocess the given data. The computation of persistence is fed by the time series data embedded in higher dimensions and provided as a distance matrix to the algorithm that computes the persistent features encoded into (persistence) diagrams. The images in Figure 2 show the cloud of input data points, the corresponding simplicial complex (a Vietoris-Rips complex), and the corresponding persistence diagram for dimension 1 (where the information on topological cycles is captured). The computation of the persistence diagrams is done using Ripser [1], an open source persistent homology software that can output a text file with a list of birth and death times corresponding to the measure of persistence, fully describing the persistence diagram. We also used an alternative open source tool, Perseus [6], whenever we needed to control parameters of persistence computations (eg. step size, number of steps or initial threshold distance). 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 Figure 3 we can see the 3-step construction of the Vietoris-Rips complex that will provide us with the persistence diagram encoding the topological information of the Influenzanet data. These topological tools complement the information obtained by classical data analysis.

III. COMPARING ILI SEASONS

When comparing two time series that may vary in time or speed it is reasonable to apply DTW to measure the similarity between the temporal sequences. Doing so, we are able to align the time series to enable comparison between seasons. In this study we compared each pair (country, year), obtaining the respective measures in the table of Figure 5, with highlighted largest and smallest values. We compare these values with those also in Figure 5 coming from comparing persistence diagrams each of which corresponding to an influenza season in either Portugal or Italy (for the embedding using and comparing Euclidean and Mahalanobis metrics). Bottleneck distance is a standard technique of TDA that permit us to measure the pairwise distance between persistence diagrams at each dimension. The distance value between two persistence diagrams in the tables of Figure 5 was calculated using the persistence landscapes toolbox [4] to compute the distance between diagrams. Persistence Landscapes generalizes



SW Persistence

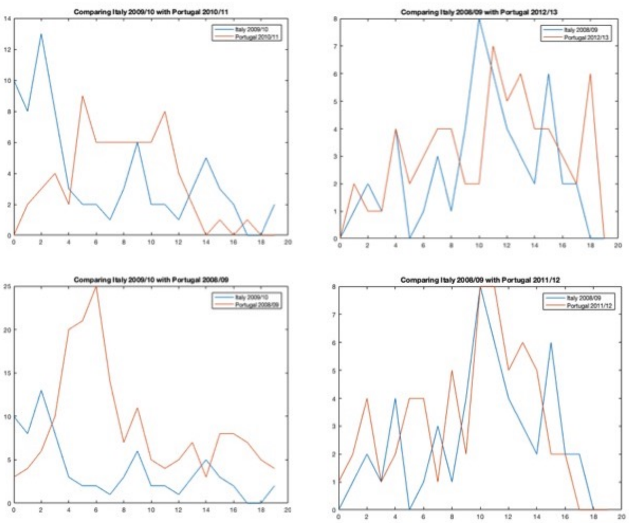
Mahalanobis

		Italy				
		2008	2009	2010	2011	2012
Portugal	2008	0.69054	0.67536	0.51681	0.66377	0.58568
	2009	0.53593	0.52165	0.37944	0.35339	0.38572
	2010	0	**0.031433**	0.24607	0.27187	0.1758
	2011	0.1758	0.18699	0.28006	0.32567	0.27187
	2012	**1**	0.98653	0.81235	0.90479	0.86585

DTW

		Italy				
		2008	2009	2010	2011	2012
Portugal	2008	0.87255	**1**	0.69608	0.57843	0.40196
	2009	0.78431	0.31373	0.73529	0.72549	0.91176
	2010	0.19608	0.47059	0.13725	0.14706	0.42157
	2011	**0.17647**	0.47059	0.2451	0.27451	0.5098
	2012	0.22549	0.54902	0.26471	0.38235	0.53922

Fig. 5. Comparing the ILI seasons of Portugal and Italy during 2008/09 up to 2012/13: the distance tables for the TDA with Euclidean metrics (on the top), the TDA with Mahalanobis metrics (on the center), and the dynamic time warping (on the bottom).

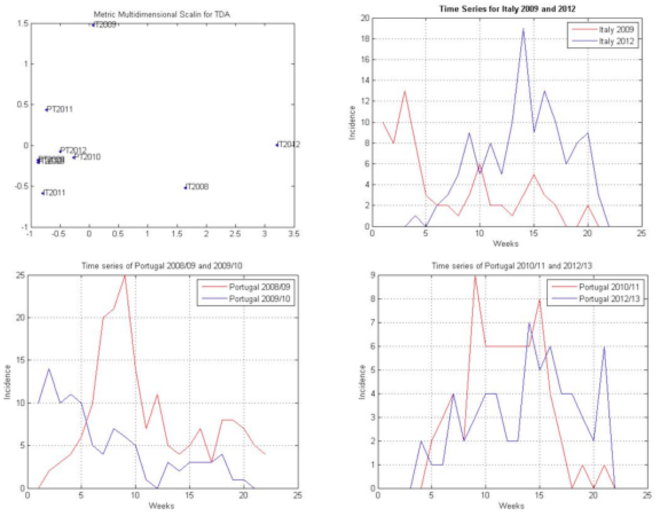

Comparing Italy 2009/10 with Portugal 2010/11

Comparing Italy 2008/09 with Portugal 2012/13

Comparing Italy 2009/10 with Portugal 2008/09

Comparing Italy 2008/09 with Portugal 2011/12

Fig. 6. Comparing the ILI seasons of Portugal and Italy during 2008/09-2012/13: selected plots of time series to compare the results in the topological data analysis and the dynamical time warping.


Metric Multidimensional Scalin for TDA

Time Series for Italy 2009 and 2012

Time series of Portugal 2008/09 and 2009/10

Time series of Portugal 2010/11 and 2012/13

Fig. 7. Comparing the ILI seasons of Italy and Portugal during 2008-2013 using metric multidimensional scaling (on the upper left) to identify: the outlier ILI seasons of Italy 2009/10 and 2012/13, with time series plotted for analysis and interpretation (on the upper right); the close ILI seasons of Portugal 2008/09 and 2009/10 (on the lower left); and the ILI seasons of Portugal 2010/11 and 2012/13, close to the diagonal (on the lower right).

bottleneck distance and will be used in further research to get deeper insights from these comparisons. The tables in Figure 5 represent the comparison between the TDA and DTW analyses of ILI incidence in Italy and Portugal for the ILI seasons of 2008/09 up to 2012/13. This data was normalized by maximum distance (i.e., using normalized = $(x - min(x))/(max(x) - min(x))$), enabling the comparison of $[0, 1]$ values in such non homogeneous data. When comparing the distances obtained by DTW and TDA we can see that these two methods look at different features of the data and thus the different results obtained. The plots in Figure 6 represent time series for the selected ILI seasons. They allow us to compare the different data analyses methods used in this study. When comparing the distances between the ILI seasons in Italy and Portugal, the TDA often disagrees with DTW (see Figure 6). The closest ILI seasons according to TDA are those of Italy 2009/10 and Portugal 2010/11, where the TDA has a lowest value of 0.0314 (due to the higher similarity of peaks) while the DTW has an average value of 0.4706. The closest ILI seasons according to DTW are those of Italy 2008/09 and Portugal 2011/12, where the DTW has a lowest value of 0.1765 (describing the similar behavior of the curves) while the TDA analysis has also a low value of 0.1758. We used multidimensional scaling as in Figure 7 to identify outliers for each of the three methods within the ILI seasons analyzed in this study. TDA provides a qualitative analysis of the time series of ILI incidence, 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 7 describe very different ILI seasons with very different peaks. On the other hand, the ILI 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 (it is worth mentioning that these were seasons influenced by the pandemic H1N1/09 virus). 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 [15]. Here lies a strong point of TDA where it can provide relevant contribution complementing other methods. The persistence diagrams of Figure 8, correspondent to the identified ILI seasons of Italy 2009/10 and 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.

IV. LOOKING FOR PERIODICITY IN THE INFLUENZANET DATA

Fourier analysis is widely used to identify patterns in a time series. In this section we discuss how the qualitative data analysis of TDA can complement the quantitative information provided by the Fourier analysis. In Figure 3 we can see the plot of the two time series and their correspondent Fourier transform. We used the time series of ILI incidence in Portugal and Italy for the ILI seasons of 2008-2013, representing non-homogeneous data. We computed the Fourier transform for each pair of time series (country, year) to compare the ILI seasons of Portugal and Italy, as in [11], confronting the quantitative methods of the Fourier analysis with the qualitative methods of TDA. TDA can also be used to look for periodicity in Influenzanet data, following the work in [14], to identify recurrent behaviours within selected influenza seasons. Barcodes and the correspondent persistence diagrams, seen as multi-scale signatures encode the lifetime of topological features

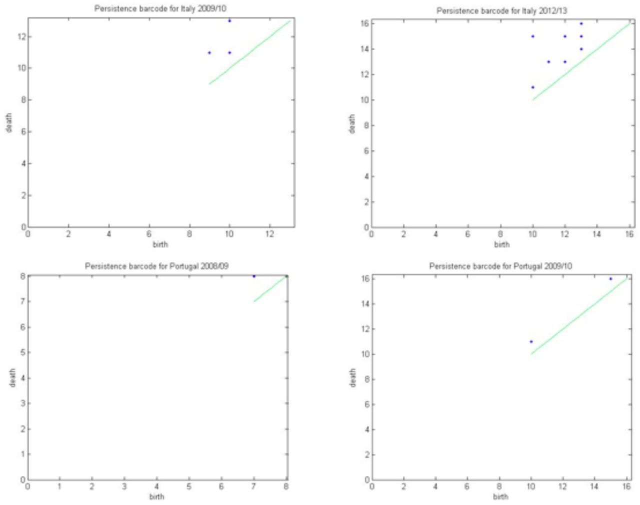


Fig. 8. Comparing the ILI seasons using persistence diagrams for dimension 1 for: Italy 2009/10 (on the upper left), Italy 2012/13 (on the upper right), Portugal 2008/09 (on the lower left), and Portugal 2009/10 (on the lower right), identified as particular cases in Figure 7.

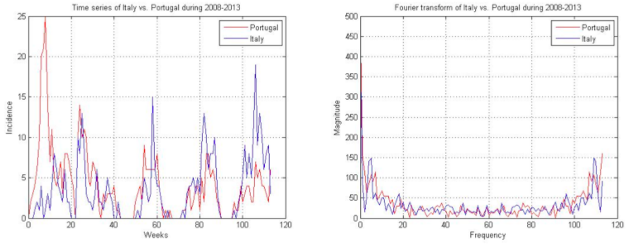


Fig. 9. Comparing the ILI seasons of Portugal and Italy during 2008-2013: the time series (on the left); the Fourier transform (on the right).

within pairs of numbers representing birth and death times. We have computed a persistence diagram in Figure 4 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.

V. 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 could be used to analyze and compare ILI seasons between countries based on the curves of the time series of their ILI incidence. The analyzed Influenzanet data lists the number of active participants and the number of ILI onsets, for three different ILI case definitions. Using the described methods we shall also look at those different ILI case definitions, contributing to a better understanding of the features distinguished by them. The information provided by quantitative methods such as DTW or the Fourier analysis of time series can be combined and complemented by the TDA analysis of that data. Further research considers the analysis of the impact of the TDA analysis for modeling and prediction of the current Influenza season. We can also complement the approach with other machine learning methods to learn metrics that are more appropriate to the input time series data, aiming to grasp a better understanding of the severity of the epidemics both in past and ongoing ILI seasons.

ACKNOWLEDGMENTS

The first and third authors would like to thank the support of Croatian Science Foundation's funding of the project EVOSOFT (UIP-2014-09-7945), and the H2020 project MIDAS (agreement 727721). The second author thanks the support of ARRS ref. TOPREP (ARRS N1-0058). A special thanks to Gabriela Gomes, co-founder of Influenzanet, for introducing the authors to this participatory monitoring system, for making the data available, and for commenting on a previous version of this work. We are also grateful to Ana O. Franco for information complementing the available data, and for revising a previous version of this work.

REFERENCES

[1] U. Bauer (2015). Ripser. *github.com/ripser*.

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

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

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

[5] L. Li et al (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine* **7.311**: 311ra174–311ra174.

[6] V. Nanda (2014). Perseus. *sas.upenn.edu/vnanda/perseus*.

[7] M. Nicolau et al (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences* **108.17**: 7265–7270.

[8] 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.

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

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

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

[12] J. Pita Costa (2017). Topological data analysis and applications. In *40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2017)*, IEEE: 558–563.

[13] Sander P. van Noort et al (2007). Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe. *Eurosurveillance* **12.7**: E5–6.

[14] Vin de Silva, P. Skraba, and M. Vejdemo-Johansson (2012). Topological analysis of recurrent systems. In *Workshop on Algebraic Topology and Machine Learning, NIPS 2012*.

[15] E. YomTov et al (2015). Estimating the Secondary Attack Rate and Serial Interval of Influenzalike Illnesses using Social Media. *Influenza and other respiratory viruses* **9.4**: 191–199.