The Euler characteristic as a topological marker for outbreaks in vector-borne disease

Danillo Barros de Souza*

Mathematical, Computational and Experimental Neuroscience team, Basque Centre for Applied Mathematics (BCAM), 48009 Bilbao, Basque Country, Spain and Departamento de Matemática, Universidade Federal de Pernambuco (UFPE), Recife, 50670-901, Brazil.

Everlon Figueirôa dos Santos

Departamento de Matemática, Universidade Federal de Pernambuco (UFPE), Recife, 50670-901, Brazil.

Fernando A. N. Santos

Departamento de Matemática, Universidade Federal de Pernambuco (UFPE), Recife, 50670-901, Brazil, Dutch Institute for Emergent Phenomena (DIEP), Institute for Advanced Studies, University of Amsterdam, Oude Turfmarkt 147, 1012 GC, Amsterdam, The Netherlands and Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 105-107, 1098 XG Amsterdam, the Netherlands

(Dated: December 1, 2022)

Abstract

Epidemic outbreaks represent a significant concern for the current state of global health, particularly in Brazil, the epicentre of several vector-borne disease outbreaks and where epidemic control is still a challenge for the scientific community. Data science techniques applied to epidemics are usually made via standard statistical and modelling approaches, which do not always lead to reliable predictions, especially when the data lacks a piece of reliable surveillance information needed for precise parameter estimation. In particular, Dengue outbreaks reported over the past years raise concerns for global health care, and thus novel data-driven methods are necessary to predict the emergence of outbreaks. In this work, we propose a parameter-free approach based on geometric and topological techniques, which extracts geometrical and topological invariants as opposed to statistical summaries used in established methods. Specifically, our procedure generates a time-varying network from a time-series of new epidemic cases based on synthetic time-series and real Dengue data across several districts of Recife, the fourth-largest urban area in Brazil. Subsequently, we use the Euler characteristic (EC) to extract key topological invariant of the epidemic time-varying network and we finally compared the results with the effective reproduction number (R_t) for each data set. Our results unveil a strong correlation between epidemic outbreaks and the EC. In fact, sudden changes in the EC curve preceding and/or during an epidemic period emerge as a warning sign for an outbreak In the synthetic data, the EC transitions occur close to the periods of epidemic transitions, which is also corroborated. In the real Dengue data, where data is intrinsically noise, the EC seems to show a better sign-to-noise ratio once compared to R_t. In analogy with later studies on noisy data by using EC in positron emission tomography (PET) scans, the EC estimates the number of regions with high connectivity in the epidemic network and thus has potential to be a signature of the emergence of an epidemic state. Our results open the door to the development of alternative/complementary topological and geometrical data-driven methods to characterise vector-borne disease outbreaks, specially when the conventional epidemic surveillance methods are not effective in a scenario of extreme noise and lack of robustness in the data.

Keywords: Euler characteristic, TDA, vector-borne diseases, complex networks

*Electronic address: danillo.dbs16@gmail.com

I. INTRODUCTION

The development of data-driven surveillance tools for epidemic outbreaks is paramount for saving human lives. There is an indispensable need for such tools to ensure the stability of public health; a case in point is the recent SARS-COV-2 outbreak [1, 2]. However, the development of such tools poses significant medical, scientific, computational, mathematical and technological challenges due to the complexity of spatial-temporal interactions of agents (humans, pathogens) and high-dimensional latent variables in the real world. The development of such tools is a hot topic in epidemiology, and various approaches have been proposed [3–6]. Early developments in the 1950s were based on time-varying models together with statistical and stochastic techniques that aimed at dissecting the inherent complexity of outbreak events. A pioneering example was the Kermack and McKendrick theory [7, 8]. Over the past years, research in complex network theory made it possible to deal with epidemic behaviour from a theoretical, statistical and big data perspective. This has, for instance, enabled sophisticated network data visualization to present population dynamics behavior and parameter estimation of mathematical models to predict epidemiological developments [9–18].

Topological Data Analysis (TDA) and Geometrical Data Analysis (GDA) are recent promising data-analytic techniques that allow extraction of non-trivial topological and geometrical features from multi-scale high-dimensional data [19–23]. This conceptual idea of analysing data via geometrical and topological means stands in stark contrast with conventional numerical summaries (e.g. used in statistics), which might require advanced inference methods and a reliable data set for a precise parameter estimation. In analogy to a piece of music (i.e. data), which is composed of chords, topological invariants can be understood as the fundamental patterns (i.e. chords) that best describe the shape of the data. The effectiveness of TDA and GDA has been shown in several applications across several fields [24–28]. For example, Ricci curvature discretization has been useful to predict fragility and risk in stock exchange [29] and cancer diagnostics [30, 31], while Betti numbers and the Euler characteristic have been applied to brain activity [32]. Given this past work, we envisage that TDA and GDA methods have potential for becoming alternative surveillance tools for epidemics. As a paradigmatic example, the authors recently developed geometric approaches in an attempt to infer early signs of COVID-19 new waves by using discrete

Ricci curvatures [33]. In line with this approach, the aim of the present work is to explore potential develop surveillance tools based on TDA as an alternative topological approach to tracking the emergence of epidemic outbreaks. Here, we will focus on analyzing data from time-series new cases of Dengue disease outbreaks.

Dengue is a vector-borne viral disease that represents a significant health concern for our globalized world, where epidemic control is still a challenge for science and society. Furthermore, its unpredictable spreading dynamics have been threatening humanity, particularly in tropical countries worldwide [34, 35]. Classical dengue fever, also known as break-bone fever, is distinguished by headache, a sudden onset of fever, sore muscles and joints, with occasional nausea and rash. These symptoms may last for several days. When it comes to Dengue hemorrhagic fever, the symptoms also include a sudden onset of fever and hemorrhagic manifestations, resulting in fluid loss and a higher risk of death [36]. Unfortunately, there is a lack of robust topological and geometrical surveillance tools for monitoring and controlling Dengue disease. Also, the lack of a well-structured data collection limit the surveillance methods to basic statistics. In our case, the lack of data prevent us to perform advanced methods to compute the effective reproduction number through the methods like the ones presented in [37, 38], which might require estimation of time-delay distribution for example.

This motivates the present work, which proposes the use of TDA methods to analyze directly the topological and geometrical changes of time-series data associated with Dengue outbreaks, which are noisy and poorly collected in general [39–42]. Proceeding in this way, we can track the epidemic landscape without the use of mathematical modelling and parameter estimation from data. In particular, the Euler characteristic emerges as an effective tool to highlight topological information from noisy data in complex networks. In [43], the authors computed the Euler characteristic of positron emission tomography (PET) scan data and concluded that the Euler characteristic is an estimator of the number of isolated activation clusters in the brain. In analogy, for an epidemic network, we hypothesise that the Euler characteristic can be interpreted as an estimate of the number of epidemic clusters and, therefore, could be a topological marker for an epidemic outbreak. Indeed, our analysis shows that, in analogy with [29, 33], there is a strong correlation between topological and epidemic curves. More specifically, the Euler characteristic could serve as a marker for Dengue outbreaks despite high levels of noise associated with poor data collection or constant

endemic situations.

This work is written as follows: In section II we present the methods and materials used in our approach, followed by section III, where we first introduce the model in which we generated our synthetic Dengue data. The results of our approach for both synthetic and real Dengue data are shown in and compared with the results of the effective reproduction number in section IV. Finally, in section V, we present the conclusion and further considerations of our work.

II. MATERIALS AND METHODS

In this section, we discuss the methods for creating time-evolving networks from epidemic data, as well as the filtration of networks and the calculus of Euler characteristics.

A. Building a network from epidemic time-series data

Network-based approaches in epidemiology are usually performed by associating nodes to individuals, while the edges represent the transmission of the disease [9, 11, 44, 45]. The starting point of our approach for creating an epidemic network differs from the standard one in that it is based on Pearson correlation coefficients computed from time-series data, and is often used as a similarity measure for discrete time-series in general [46]. This approach is inspired by network analysis in other fields, such as neuroscience [47] or finance [48], where the connectivity between two nodes is often functional and not only structural.

Guided by the work in [33], we define the evolving epidemic networks from time-series in this section. More specifically, we define the geographical places where the new cases of epidemic time-series are collected as the nodes of the network and the (weighted) links are pairwise Pearson correlation coefficients between the provided time-series [49–51]. As a result, we have a (weighted) undirected network carrying the epidemic information from the pairwise Pearson correlation between these time-series. Naturally, this network can be translated into a symmetric square adjacency matrix, where the entries are the correlation coefficients between the nodes.

Formally, for every fixed time interval in the epidemic data, we generate an undirected simple weighted graph $G = (V, E)$, where V are the places that provide the time-series and

$E = \{e = (x, y);\, x, y \in V\,,\, x \neq y\}$ is the set of edges, with $w_e \in [-1, 1]$ the weight of the edge $e = (x, y)$, computed according to the Pearson correlation coefficient between the places x and y. By repeating this step for all time intervals, we build a sequence of time-evolving networks that carries the information of the epidemic data along the time.

B. Filtration method

The procedure of generating networks from data generally results in dense graphs, which may have several spurious links or might be time demanding for most large-scale complex networks analyses. Also, there is no established way to threshold a network and, therefore, no established filtration procedure. Taking these facts into account, we decided to use a network filtration method that allows us to delete spurious information from the network. This process is also performed in our recent work [33] and is described as follows:

Let $G = (V, E)$ be an undirected weighted simple graph. The *ϵ-neighbourhood* [19] is a subset of E that derives a subgraph $G_\epsilon \subset G$, defined by $G_\epsilon = (V, E_\epsilon)$, and is defined as

$$E_\epsilon = \{e = (x, y) \in E, x \neq y\,;\, m(x, y) \leq \epsilon\}, \tag{1}$$

were $m(x, y)$ is a metric distance defined over the set of nodes. The added value of this definition is the possibility of visualizing an evolving graph as a function of a parameter ϵ.

Note that the interaction between nodes is provided by the Pearson correlation coefficient matrix, whose entries might assume negative values and cannot be considered as distance values in (1). In other domains, the absolute value of the Pearson correlation coefficient is often considered to define these networks. However, the original information is useful for measuring the degree of synchronicity of the time-series: for example, a negative correlation between two time-series can be interpreted as opposite trends of epidemic stages - the number of new epidemic cases increases in one place while it decreases elsewhere. Furthermore, the idea of varying the ϵ values is to start with an empty graph and gradually add the *strongest* links to the evolving graph, i.e., we first add the edges with a higher correlation coefficient. This construction can be harmed by considering the absolute value of Person correlation coefficients as done in other domains.

For these reasons, we remodeled eq. (1) by setting E_ϵ as

$$E_\epsilon = \{e \in E;\ w_e \geq 1 - \epsilon\}, \tag{2}$$

where w_e is the weight of the edge e. From (2), we can see that ϵ runs over the interval $[0, 2]$ (once the Pearson correlation values run in $[-1, 1]$) and that $G_\epsilon \subset G$. In order to ignore redundant information for the time-varying analysis, we compute the critical percolation value of ϵ_c such that the graph still keeps the connections which are relevant in the skeleton structure. This threshold value is defined as follows:

$$\epsilon_c = \inf\{\epsilon \in [0, 2]\ ;\ |G_\epsilon| = |G|\}, \tag{3}$$

where $|G|$ denotes the number of connected components of $G = G_2$ [52, 53]. The idea is to keep the graph structure the same as the crossing number of connected components (or, alternatively, the *Betti numbers* of G [54]). The filtration process is visualized in Figure 1. In this example, the original network (ϵ=2.0) has the same number of connected components as its filtered version ($\epsilon_c = 0.63$), which is the chosen threshold network that preserves the number of connected components of the original network. For each time-evolving graph,

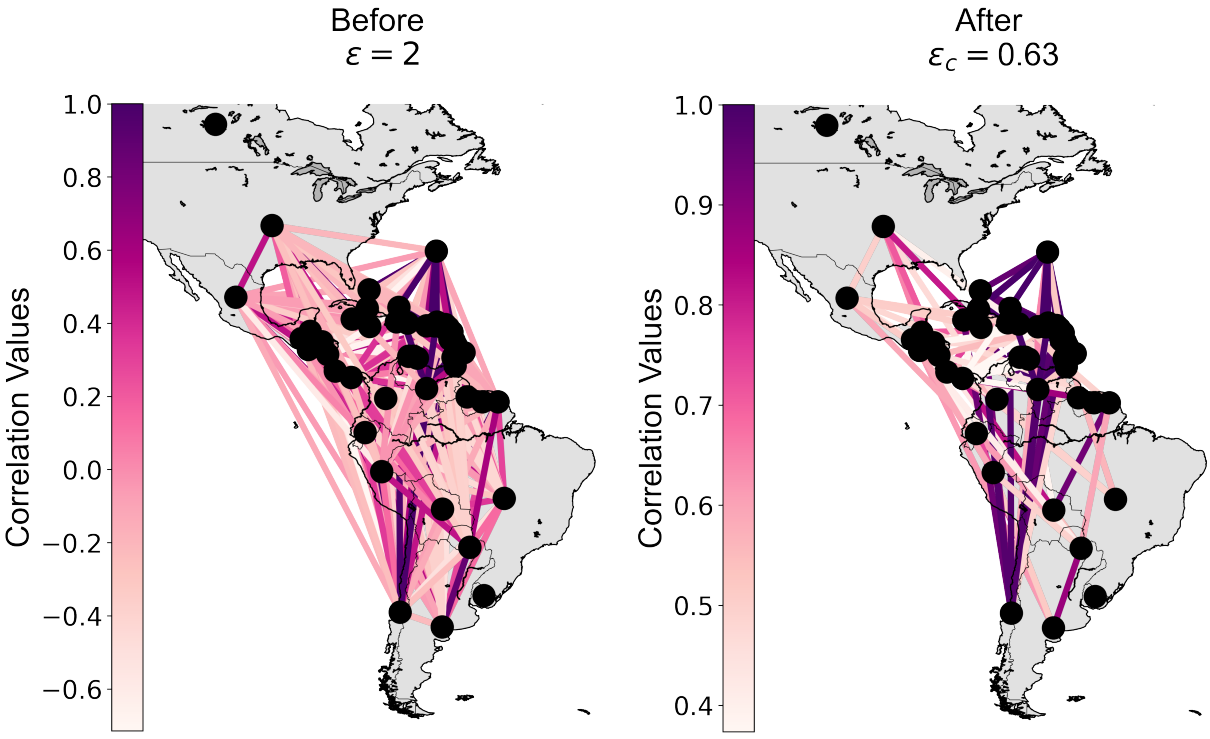

FIG. 1: Visual and geometrical difference between non-filtered network (left) and filtered network (right).

a filtration is selected in the vicinity of the giant component transition. As is discussed in theoretical models for epidemic networks, an epidemic outbreak happens at the critical probability for the emergence of a giant component transition [11, 13], so that the detachment of connected components is imminent. Here, we are seeking the smallest threshold value that leads to the maximum amount of connected components in the graph. This approach not only reduces time-processing and spurious links but also maintains the crucial information provided by the original network. In the following subsections, we will recall the basics of CW complexes relevant for our analysis and their applicability to computing the Euler characteristic of complex networks.

C. CW complexes

The study of topology in a continuous manifold and its geometric counterparts is not a new field [55–59]. However, discrete versions of theoretical results in differential geometry and associated computational algorithms have recently become a burgeoning area of pure and applied mathematics [60]. Over the past years, the topological and geometric approaches were formalized for discrete structures [56, 61–65]. Here, we discuss the relevant topological objects for our analysis and, in particular, introduce the Euler characteristic as a discrete topological measure that can be written in terms of cells of a CW-complex.

We are going to introduce the concept of CW-complexes and how this structure relates to the computation of Euler characteristic.

Let $G = (V, E)$ be an undirected graph. We define a d-cell as a complete subgraph of G with $d+1$ nodes, *i.e.*, the structures equivalent to open balls of dimension d in the continuous definition [66]. They are also called $(d+1)$-vertex cliques (or cliques of dimension d) in the computational approach. The union of all d-cells is said a *CW-complex*.

In Figure 2, we see a CW-complex consisting of 10 cells of dimension 0 (nodes), 13 cells of dimension 1 (edges), 5 cells of dimension 2 (triangles) and 1 cell of dimension 3 (tetrahedrons). Alternatively, we can also say that the structures in this figure are cells of dimensions $0, 1, 2$ and 3, respectively.

In our work, the CW-complexes are extracted from the time-evolving filtered networks defined in the last section. These structures are relevant for expressing the topological signature of the epidemic data, in this work, expressed in terms of Euler characteristics.

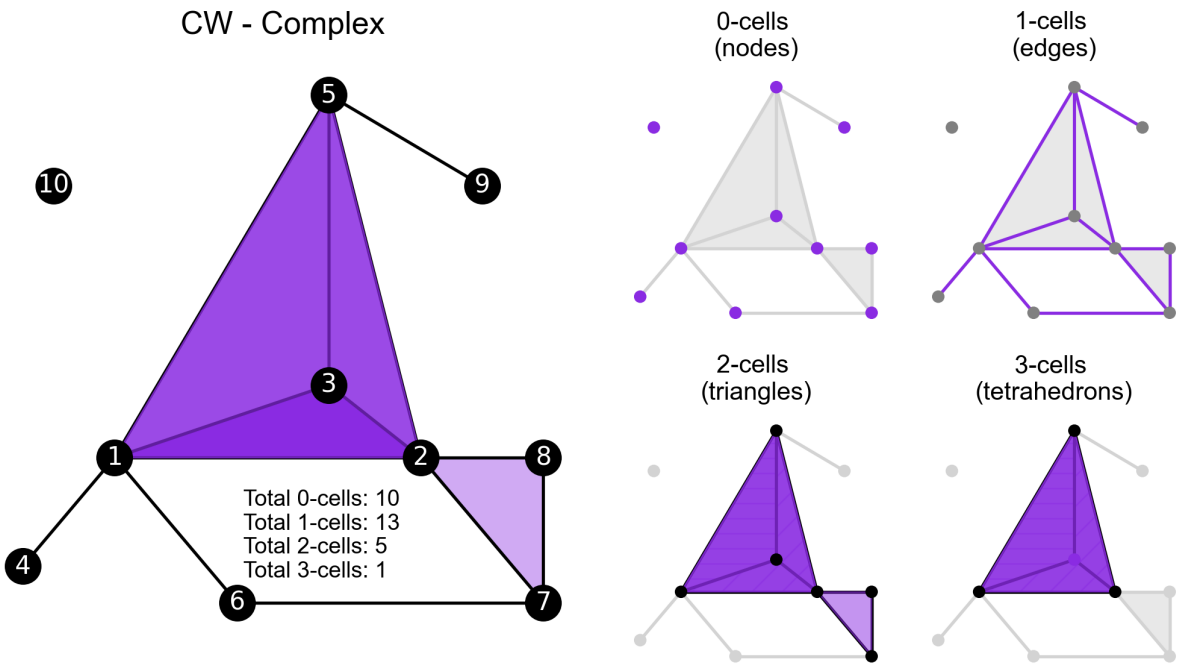

FIG. 2: Example for a CW-complex and its d-cells (or $(d+1)$-vertex cliques) highlighted in purple,for $d \in \{0, 1, 2, 3\}$.

D. The Euler characteristic

One important set of tools used to explore and understand the shape of data is Topological Data Analysis [23, 67]. Over the past years, TDA applications across fields have yielded astounding results for science. For instance, topological transitions have been used as bio-markers in protein-protein networks [68] and to identify the emergence of neurological diseases from brain data[32]. There are many ways to compute the Euler characteristic across fields, [69–73]. Here, we use [71] in order to reduce the computational complexity. The Euler characteristic of a CW-complex is defined as follows:

Let $G = (V, E)$ be an undirected finite graph. We define the *Knill curvature* of a node $v \in V$ as [65, 74, 75]

$$K(v) = 1 + \sum_{d=1}^{d_{\max}} (-1)^d \frac{S_d(v)}{d+1}, \tag{4}$$

where S_d is the number of d cells containing v and $d_{\max}$ is the highest cell (clique) dimension of the CW-complex. The d-cells are the same as defined in the previous section. Equation

Node	Knill Curvature
1	-3/4
2	-5/12
3	1/4
5	-1/4
6	0
7	-1/6
8	1/3
4	1/2
9	1/2
10	1

TABLE I: Comparison between Knill curvatures across nodes, based on the CW-complex in Figure 2.

(4) can also be re-written as follows:

$$K(v) = 1 + \sum_{\substack{c \supseteq v, \\ c \in C}} \frac{(-1)^{\#c-1}}{\#c}, \tag{5}$$

where c is a cell in the CW-complex C and $\#c$ is the number of nodes of the cell. The formulas in (4) and (5) satisfy the Gauss-Bonnet Theorem [65, 74, 76, 77] *i.e.*,

$$\chi(G) = \sum_{v \in V} K(v), \tag{6}$$

where $\chi(G)$ is the *Euler characteristic* of G. In Table I, it is possible to see how the Knill curvatures of the network in Figure 2 differ from each other across nodes. The Euler characteristic for the graph in Figure 2 is the sum of Knill curvatures per node in Table I, which is $\chi = 1$.

E. Computing the Euler characteristic from the dataset

Here, we are going to construct a pipeline to compute the Euler characteristic from epidemic time-series. We establish the conversion of time-series data files into time-evolving

graphs, which then enable us to compute the Euler characteristic from the dataset. This workflow is composed of the following steps:

- **Step 1:** A time window is selected from the time-series dataset;
- **Step 2:** A pairwise time-series correlation matrix is constructed from the Pearson correlation coefficient [49–51];
- **Step 3:** A network is built from Step 2 in a natural way, i.e. the information of weights and links in the network are provided by the rows (columns) of the matrix;
- **Step 4:** The network undergoes a filtration in order to reduce spurious links, through equation (3);
- **Step 5:** The Euler characteristic of the resulting filtered network is computed by using equation (6), and provides the information of the curvature for a selected time interval.

This process is repeated for all time windows in the epidemic data. Figure 3 visually illustrates the steps of this process.

In the next section, we are going to generate the synthetic data in order to study its topological behaviour afterwards.

III. APPLICATION OF TOPOLOGICAL APPROACHES TO SYNTHETIC DENGUE DATA

In order to test the validity of our hypothesis that the Euler characteristic is a topological marker for a vector-borne disease outbreak in epidemic time-series, we first apply the provided methods to synthetic epidemic data as a proof of concept. For this, we simulated time-series from a spatial epidemic interaction based on a classical stochastic epidemic disease network model. More specifically, we used a variation of a stochastic Susceptible-Infected-Recovered (SIR) model in order to replicate the behaviour of Dengue disease [78], originally used for modelling the Dengue epidemic in Rio de Janeiro, Brazil. Our approach includes birth and death interactions for generating epidemic waves. We model the infection dynamics of the population as follows:

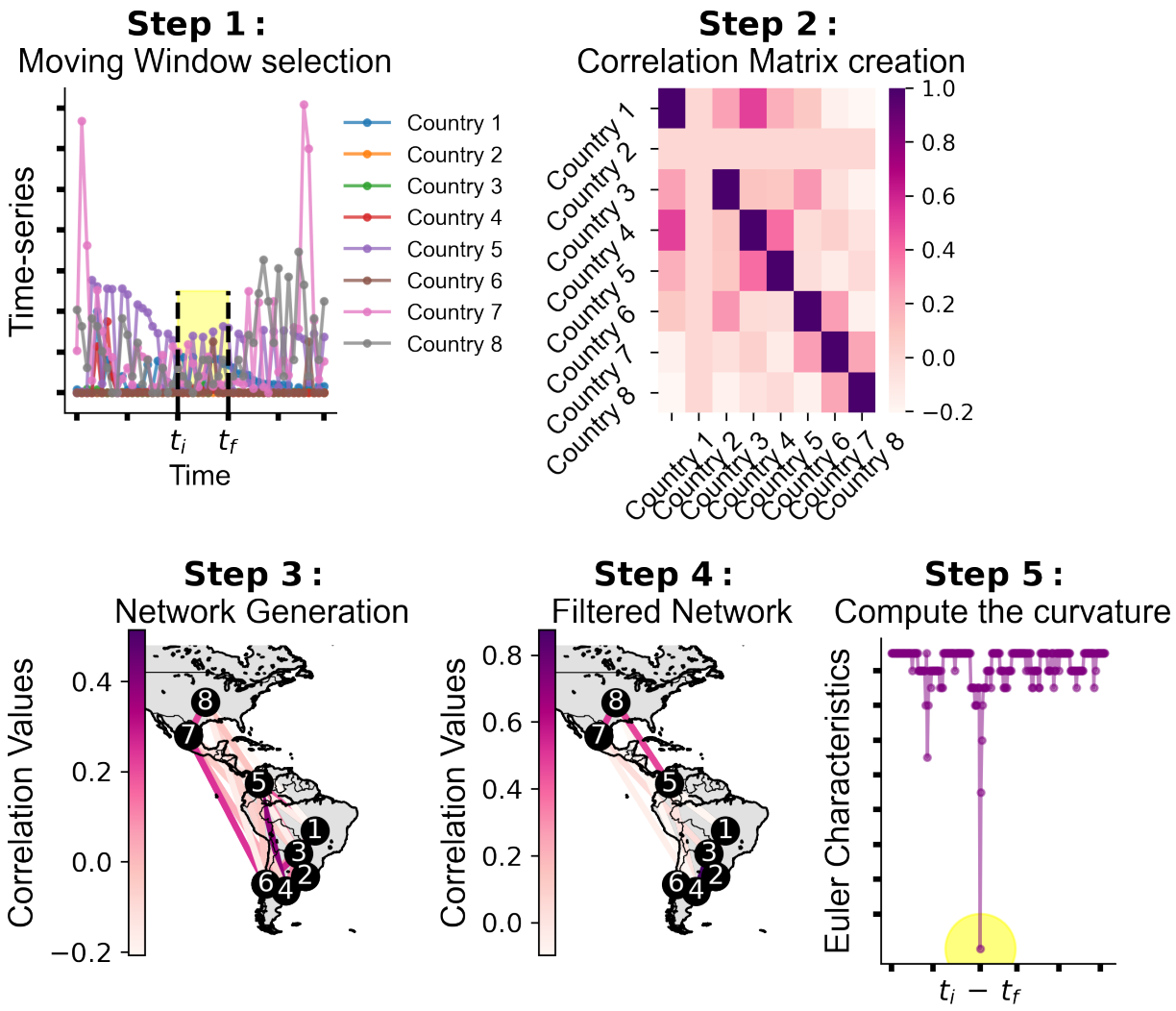

FIG. 3: An explanation of the pipeline for computing the Euler characteristic from epidemic data.

The total initial population (N) is divided by sites/places ($P_1, P_2, \cdots, P_m$). For each $j \in \{1, \cdots, m\}$ we have susceptible (S_j), infected (I_j) and recovered/removed (R_j) as local population.

The contagion dynamics between the sites is ruled by a probability matrix, $\Phi = (\phi_{ij})_{m\times m}$. Each class (susceptible, infected or recovered) has a *death rate* (δ_{S_j},δ_{I_j} and δ_{R_j}), and only the class of susceptibles has a *birth rate* α_i, for all $j \in \{1, \cdots, m\}$.

From this dynamic, we can express a system of ordinary differential equations from rate

equation techniques provided in [79],

$$\begin{cases} \frac{dS_i}{dt} &= \alpha_i - \sum_{j,k=1}^{m} \beta_j \phi_{ij} \phi_{kj} S_i \frac{I_k}{N_j^p} - \delta_{S_i} S_i \\ \\ \frac{dI_i}{dt} &= \sum_{j,k=1}^{m} \beta_j \phi_{ij} \phi_{kj} S_i \frac{I_k}{N_j^p} - (\gamma_i + \delta_{I_i}) I_i \quad , \\ \\ \frac{dR_i}{dt} &= \gamma_i I_i - \delta_{R_i} R_i \end{cases} \tag{7}$$

for all $j, k \in \{1, \cdots, m\}$, and satisfying the constant initial condition equation

$$S_j(0) + I_j(0) + R_j(0) = N_j \, , \tag{8}$$

for $N_j \in \mathbb{N}^+$, where $\sum_{j=1}^{m} N_j = N$, $N_j^p = \sum_{k=1}^{m} \phi_{kj} N_k$, and ϕ_{ij} is an element of the *Flux Matrix* satisfying

$$\sum_{k=1}^{m} \phi_{jk} = 1, \tag{9}$$

for all $j \in \{1, \cdots m\}$. In our approach, we are considering

$$\phi_{ij}(\rho) = \begin{cases} \rho, & i \neq j \\ 1 - (m-1)\rho, & i = j \end{cases} , \tag{10}$$

for some $\rho \in [0, 1/(m-1)]$. Equation (10) represents the migration rate from place i to j. In particular, when $i = j$, we have the rate of population that are not migrating (i.e. that is staying at place i). This flow is uniform when $\rho = 1/m$. We generated the epidemic time-series from a Monte-Carlo simulation [80] as follows:

1. Select the time: t;

2. Take $r \in (0, 1)$ a random variable uniformly distributed;

3. Update the propensities *i.e.*, for all $i \in 1, \cdots, m$:

   - $a_i = \sum_{j,k=1}^{m} \beta_j \phi_{ij} \phi_{kj} S_i(t) \frac{I_k(t)}{N_j^p}$;
   - $b_i = \gamma_i I_i(t)$;
   - $c_i = S_i(t)$;
   - $d_i = \delta_{S_j} S_j(t)$;

- $e_i = \delta_{I_j} I_j(t)$;
- $f_i = \delta_{R_j} R_j(t)$;
- $r_0 = \sum_{i=1}^{m} a_i + b_i + c_i + d_i + e_i + f_i$;

4. Compute the extreme points of intervals for the next time step:

$$r_i = \begin{cases} \frac{1}{r_0} \sum_{j=1}^{i} a_j, & 1 \leq i \leq m \\ \frac{1}{r_0} (\sum_{j=1}^{m} a_j + \sum_{j=1}^{i-m} b_j), & m+1 \leq i \leq 2m \\ \frac{1}{r_0} (\sum_{j=1}^{m} (a_j + b_j) + \sum_{j=1}^{i-2m} c_j), & 2m+1 \leq i \leq 3m \\ \frac{1}{r_0} (\sum_{j=1}^{m} (a_j + b_j + c_j) + \sum_{j=1}^{i-3m} d_j), & 3m+1 \leq i \leq 4m \\ \frac{1}{r_0} (\sum_{j=1}^{m} (a_j + b_j + c_j + d_j) + \sum_{j=1}^{i-4m} e_j), & 4m+1 \leq i \leq 5m \\ \frac{1}{r_0} (\sum_{j=1}^{m(} a_j + b_j + c_j + d_j + e_j) + \sum_{j=1}^{i-5m} f_j), & 5m+1 \leq i \leq 6m \end{cases}$$

5. Update time, $t := t + 1$;

6. Update population:

   - If $0 \leq r < r_1$, then $S_1(t+1) = S_1(t) - 1$ and $I_1(t+1) = I_1(t) + 1$;

   Considering the other possible cases, when $r_i \leq r < r_{i+1}$:

   - If $i \in \{1, \cdots, m-1\}$, then $S_{i+1}(t+1) = S_{i+1}(t) - 1$ and $I_{i+1}(t+1) = I_{i+1}(t) + 1$;
   - If $i \in \{m, \cdots, 2m-1\}$, then $I_{i+1-m}(t+1) = I_{i+1-m}(t) - 1$ and $R_{i+1-m}(t+1) = R_{i+1-m}(t) + 1$;
   - If $i \in \{2m, \cdots, 3m-1\}$, then $S_{i+1-2m}(t+1) = S_{i+1-2m}(t) + 1$;
   - If $i \in \{3m, \cdots, 4m-1\}$, then $S_{i+1-3m}(t+1) = S_{i+1-3m}(t) - 1$;
   - If $i \in \{4m, \cdots, 5m-1\}$, then $I_{i+1-4m}(t+1) = I_{i+1-4m}(t) - 1$;
   - If $i \in \{5m, \cdots, 6m-1\}$, then $R_{i+1-5m}(t+1) = R_{i+1-5m}(t) - 1$;

7. Return to 1. and repeat the algorithm for the next time step.

Before moving on, it is important to note that the focus of our work is to generate a toy model for epidemic time-series for further topological analysis , rather than studying the ergodocity of the model through a topological perspective, once it might be time demanding for both generating the synthetic data and computing the Euler Characteristics. Furthermore, we are not concerned with the behaviour of the stochastic model or even with parameter estimation for curve fitting with real data.

IV. RESULTS

In what follows, we will present the results of our analysis for both synthetic and real data.

A. The Euler characteristic of synthetic epidemic Dengue data

To test the validity of our proposed algorithm (outlined in the previous sections and coded in python [81]), we first apply our methodology to synthetic epidemic time-series. To this end, we set up the SIR model with a population of size $N = 2000$, with one infected individual at time $t = 0$, and we run the simulations for a total of 5000 time steps. The population was uniformly distributed into 20 (geographical) sites ($m = 20$) and the parameters for the population dynamics were $\beta_i = 1$, $\gamma_i = 0.1$, $\alpha_i = 1$, $\delta_{S_i} = \delta_{I_i} = \delta_{R_i} = 1/60$, for all $i \in \{1, \cdots, m\}$, $\rho = 1/m = 0.05$.

The resulting simulation is shown in Figure 4. The generated epidemic data shows three outbreaks in the initial, intermediate and final period of the simulation run. The first peak is the highest, and the remaining epidemic manifestations are on smaller scales, which might be compared with the mild levels of epidemic close to endemic periods. The computation of the Euler characteristic for various maximum cell dimensions ($d_{max} \in \{2, \ldots, 10\}$) in a time window of 150 points and the comparison with the synthetic epidemic data and its effective reproduction number (R_t) computed by the Joint Research Centre (JRC) method [82] is provided in FIG. 5.

To fully clarify these results, we have prepared a detailed online dynamic dashboard [83, 84], where the reader can further scrutinise the results for other time windows (for $25, 50, 75, 100, 125$ and 150 density points). As observed, the Euler characteristic manifests higher values during an epidemic phase, in analogy to the R_t transitions. These signs of the Euler characteristic might indicate the starting and/or ending period of an epidemic (for example, in the intervals $0 - 629$, $1080 - 1229$ and $1200 - 1349$), and its intensity is also sensitive to the intensity of the epidemic waves. In general, even (odd) values of d_{max} result in positive (negative) values of the curvature. The balance between the signal and the noise can be controlled by varying the maximum cell dimension and the size of the time window. The higher the dimension, the higher the computation complexity and the smaller the noise;

however, the intensity of the sign is also reduced.

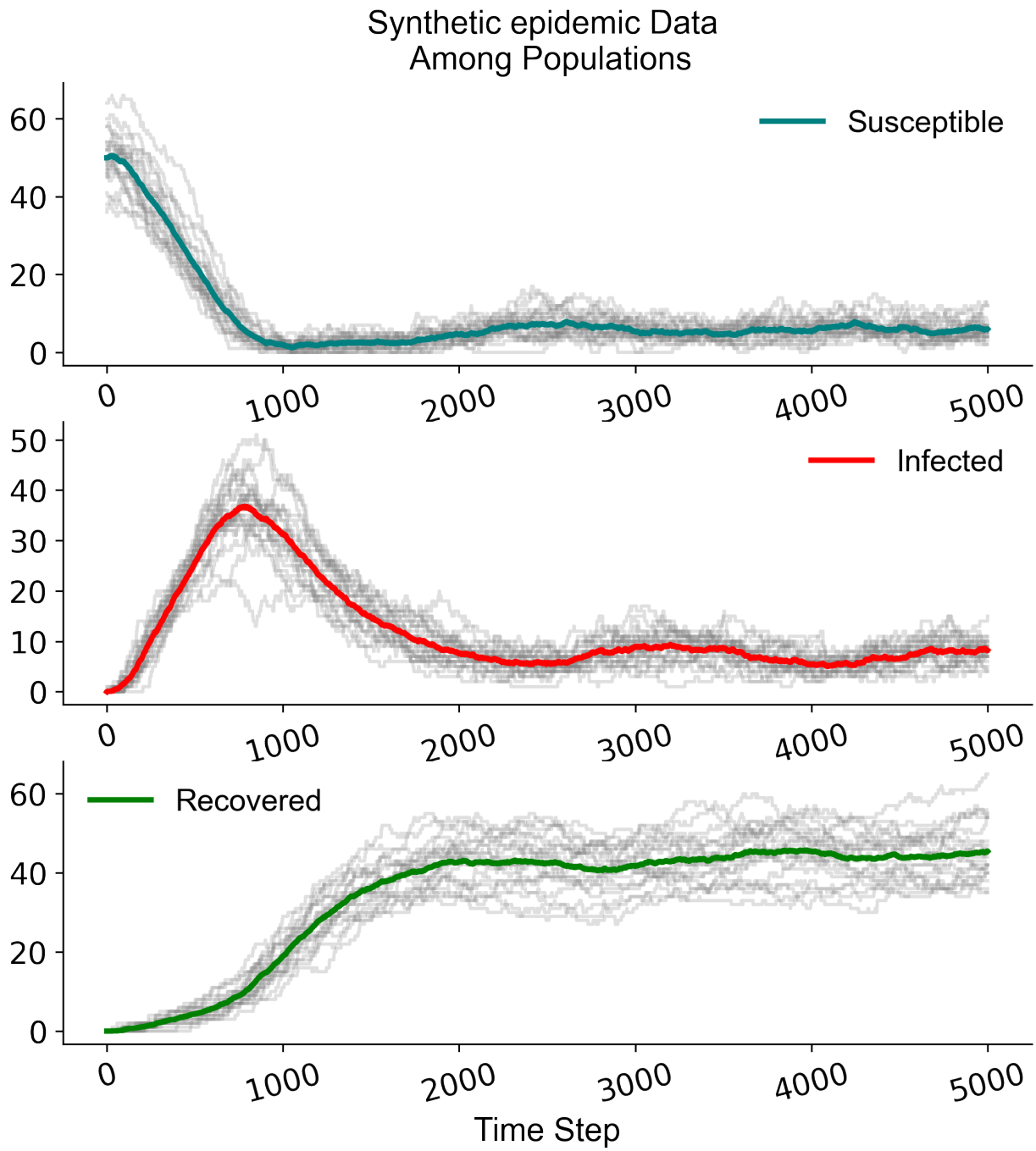

FIG. 4: Stochastic simulation of SIR epidemic model with birth and death. Each grey-coloured curve is the time-series of a site (20 sites in total). The blue, red and green curves represent the average of susceptible, infected and recovered population per time scale (in this order).

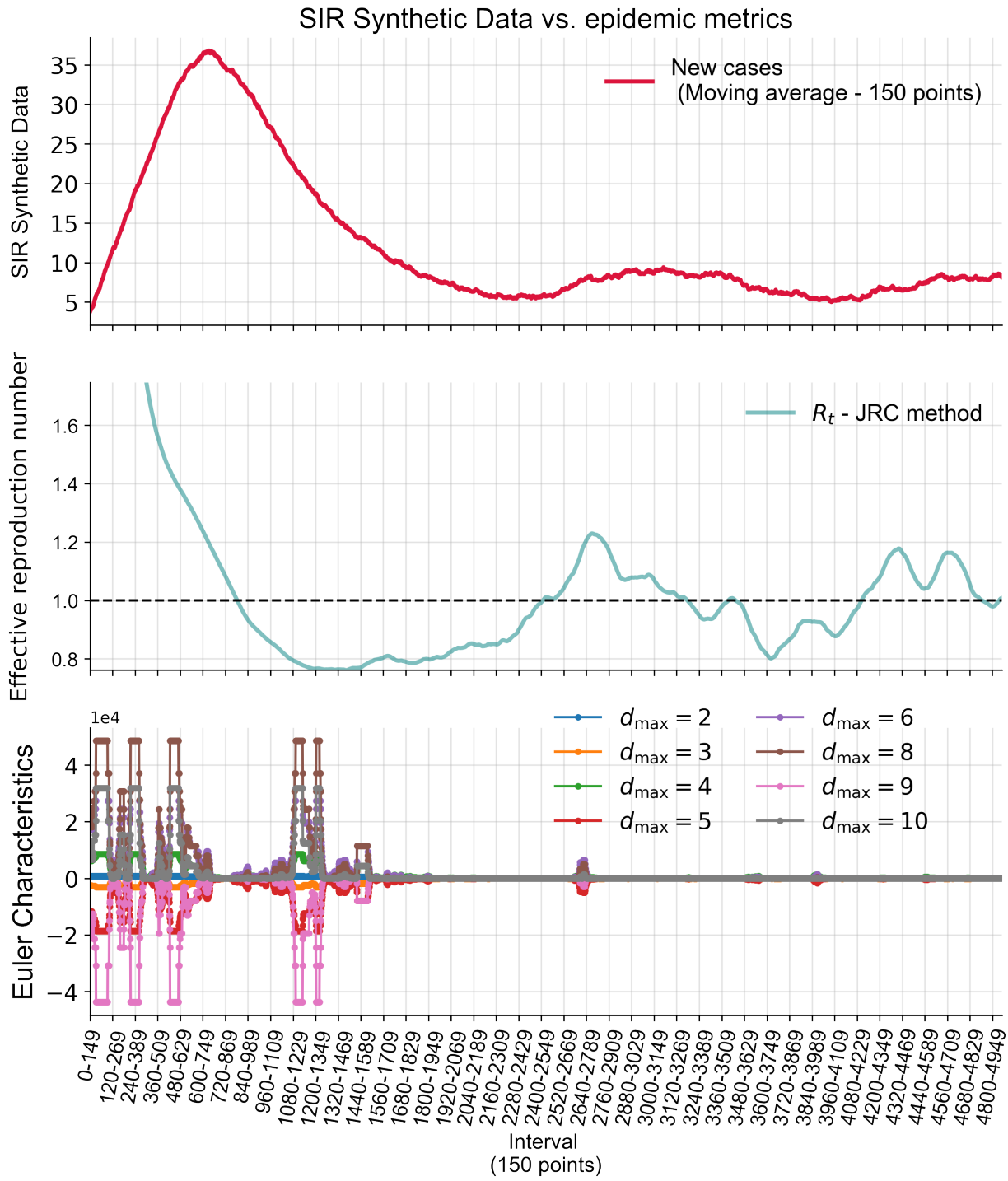

FIG. 5: Comparison between the moving average of Dengue synthetic data provided by the infected population from Figure 4 (top), the effective reproduction number (center) and its Euler characteristic (bottom) for the same time window (150 points), for various values of d_{max}.

Despite the evidence of Euler characteristic as a hallmark of the epidemic in the presented synthetic data motivated our approach, we believe that a more detailed, systematic simulation of the Euler characteristic across different epidemic models deserves further investigation.

Overall, the Euler characteristic applied to synthetic Dengue networks, together with the network filtration method, has a reasonable performance in forecasting surges in new cases, even in the presence of noise during the height of the epidemic.

B. The Euler characteristic as a possible fingerprint for epidemic outbreaks in real Dengue data

The validity of the Euler characteristic as a possible indicator for epidemic periods (as tested in our toy model) prompts us to examine it in the context of real Dengue data. In particular, we focus on Recife, Brazil's fourth largest metropolitan area in the Northeast of the country, with tropical weather and its population distributed over 94 districts/regions. Recife is a relevant city to test our hypothesis, as it experiences recurrent Dengue outbreaks over the last decades and is in a constant endemic state. The associated data can be found in the following database [85], and alternatively, we also provide all relevant information, including the raw data in our repository.

In Figure 6, we plot Recife's total new Dengue cases per day from the 1st of January 2014 to the 31st of December 2021. In FIG. 7 we compare the moving average of Dengue new cases with the R_t (JRC method [82]) and the Euler characteristics for a time window of 7 days. In Figures 8 - 9, we apply our algorithm, and we illustrate the performance of the Euler characteristic over odd and even values of d_{max}, respectively. Due to the high computational complexity, we restricted the computations for at most $d_{\max} = 7$. As in the case of the synthetic data, we also provide a detailed online dynamic dashboard [84, 86], where the reader can further scrutinise the results for other time windows ($7, 14, 21$, and 28 days).

It is noteworthy that the data shows a high level of noise due to poor public health data acquisition systems, which involve delays and other non-trivial factors. Apart from that, nothing else but the new cases time-series is provided by local health authorities in Recife (Brazil). At the same time, there is an ongoing endemic condition in the city. Despite high

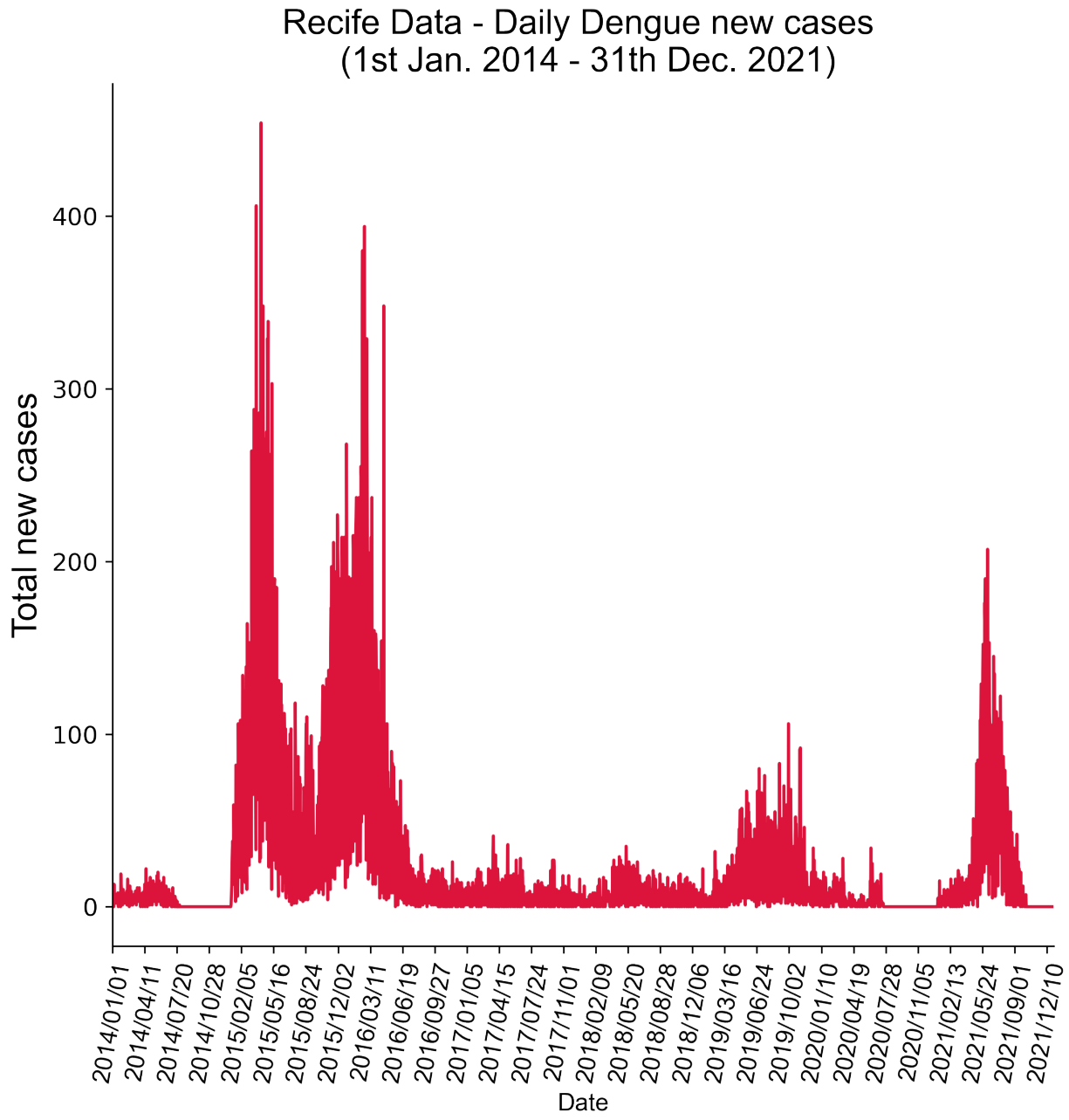

FIG. 6: Daily new cases of Dengue in Recife, Brazil.

noise levels, the Euler characteristic can increase the signal-to-noise ratio of an epidemic period. Just as in the case of the toy model, the higher $d_{\max}$, the smaller the effect of noise on the computed Euler characteristic. Moreover, even (odd) values of $d_{\max}$ lead to positive (negative) values of the Euler characteristic. This fact is explained by the exponential growth of the number of cells (cliques) at the vicinity of the outbreak, since the epidemic network becomes strongly connected.

In contrast with the results obtained in the previous section for synthetic data, the Euler characteristics correlate with the peaks of local epidemic periods rather than early warnings. On the other hand, the R_t seemed unable to catch the epidemic periods in the real data. In fact, for real epidemic time series, both facts might be explained by the delays in data collection and the generation of several zero cases in the time series. It is also essential to highlight that the peaks of R_t preceding the periods of February 2015 and February 2021 are induced by the noise generated by the zero cases in these periods.

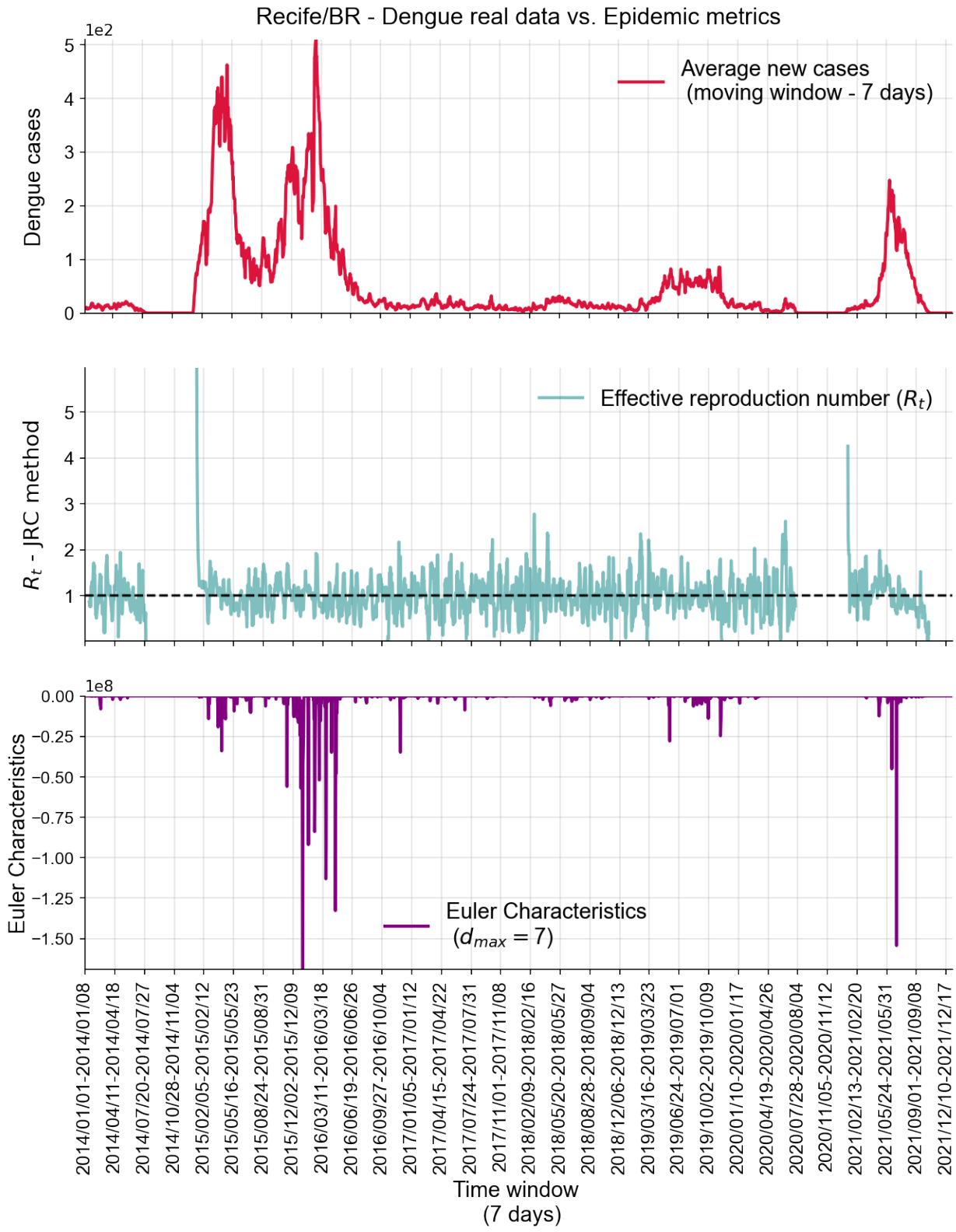

FIG. 7: Comparison between the Dengue data from Recife (red curve),its R_t and Euler characteristic, respectively.

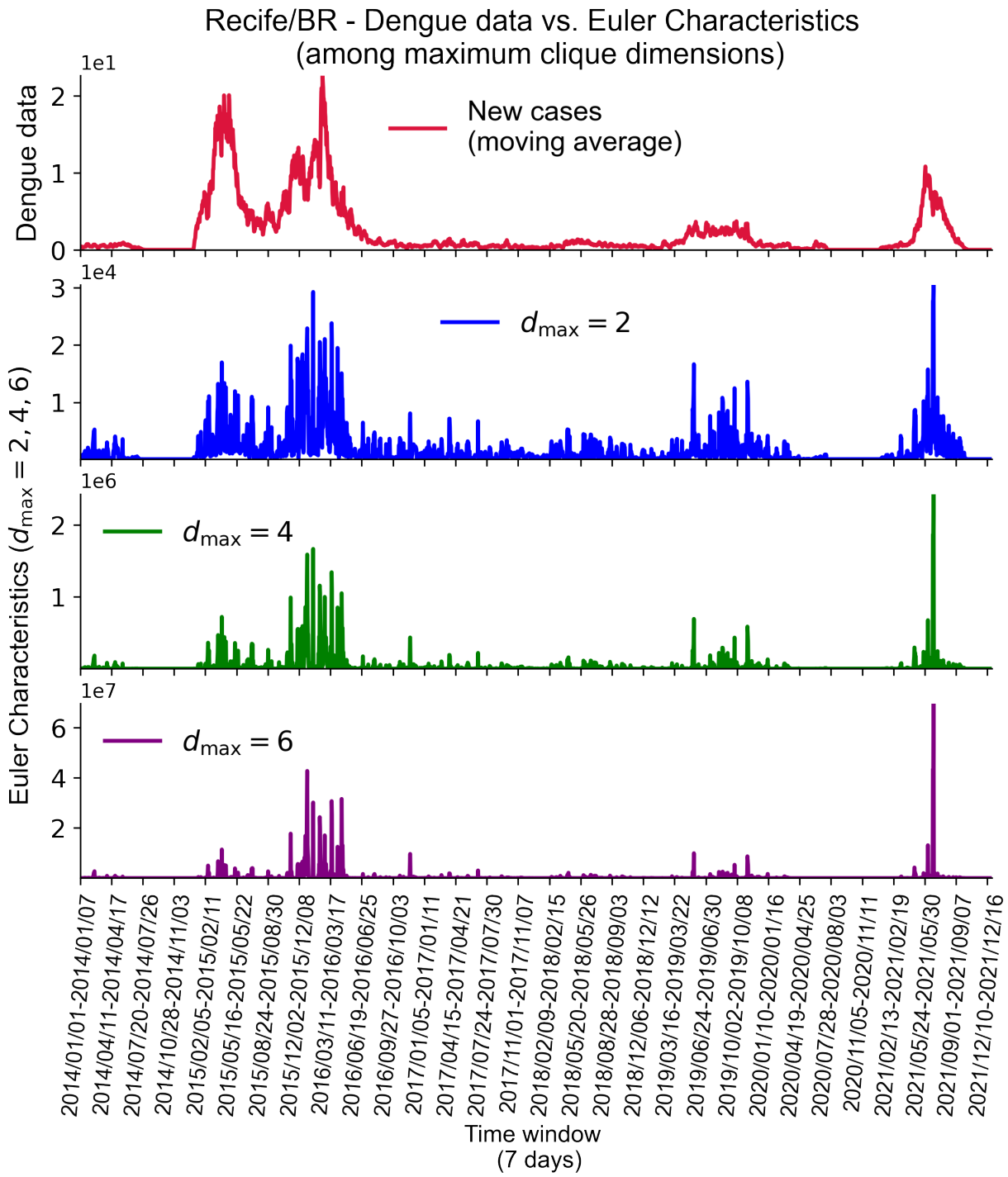

FIG. 8: Comparison between the Dengue data from Recife, Brazil, and its respective Euler characteristic, for even values of $d_{\max}$.

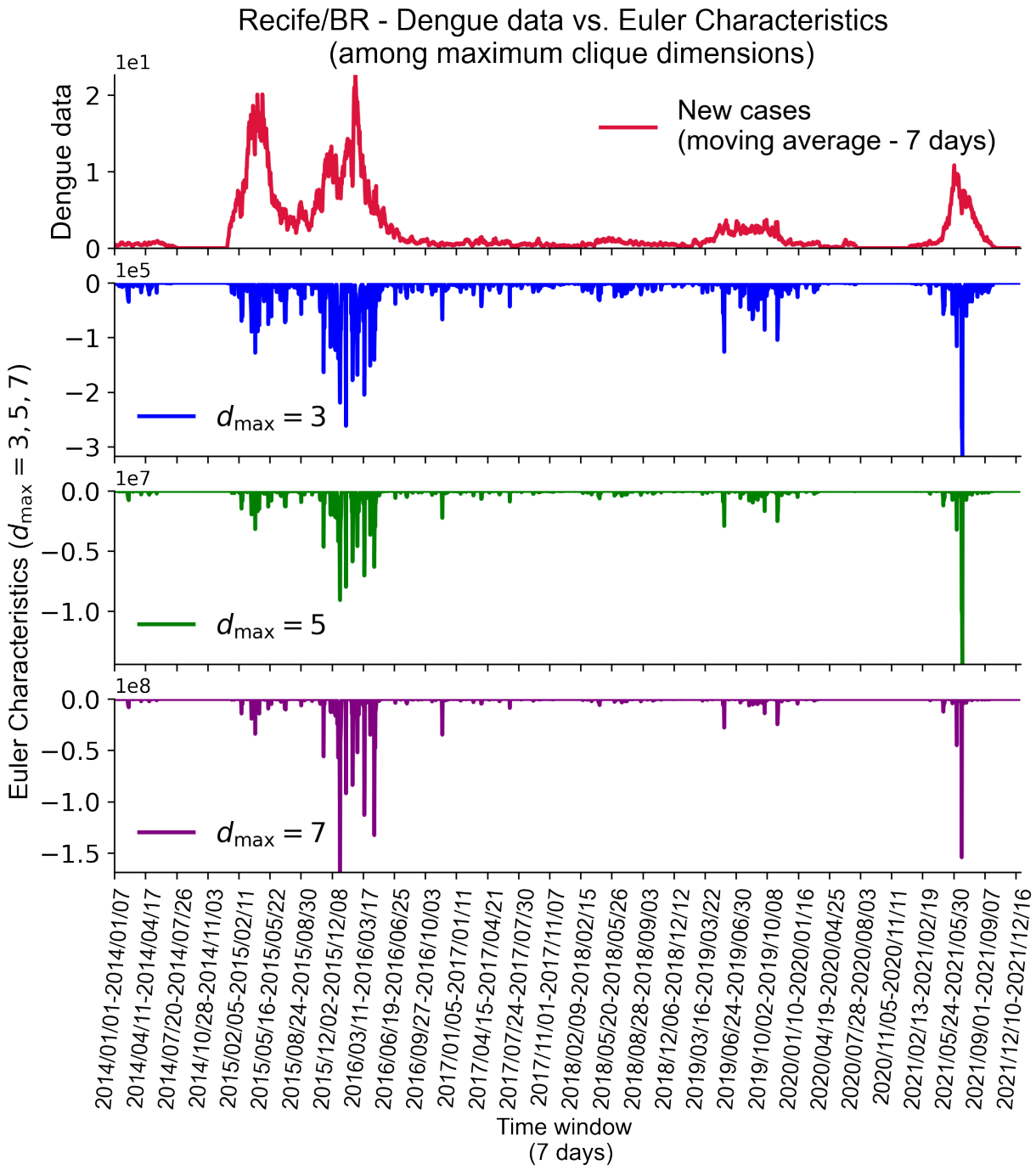

FIG. 9: Comparison between the Dengue data from Recife, Brazil, and its respective Euler characteristic, for odd values of d_{max}.

V. CONCLUSION

The present work investigates the feasibility of implementing surveillance algorithms based on topological and geometrical data analysis to infer the emergence of an outbreak in vector-borne diseases. While we focus on the example of the Dengue fever, the developed methods are general enough to be applied to other vector-borne diseases, such as Zika and Chikungunya.

Our approach was first tested on synthetic data and subsequently applied to real data. Specifically, a modified SIR stochastic model generated the synthetic epidemiological evolving network, which simulates multiple epidemic waves by different sites. Later, we computed the Euler characteristic for various time windows and cells dimensions. In particular, we showed that the Euler characteristic has significant variations during epidemic periods and more minor variations in endemic periods. Furthermore, the topological transitions of EC showed up at the vicinity of time windows in which the R_t performs its epidemic transitions for synthetic data. This indicated the viability of using topological and geometrical data analysis approaches for vector-borne diseases, which motivated our subsequent study on real data. Specifically, for the real data, we focused on the city of Recife (Brazil), which has been at the epicentre for vector-borne diseases several times and lacks of reliable data for advanced inference methods. Following our approach, we computed the evolving epidemiological network from the data on Recife and subsequently computed the Euler characteristic. In this case, the EC seems to present a better signal-to-noise ratio in its topological transitions, in comparison with the epidemic indicators of the R_t. This can be again explained by the high levels of noise in the data. We find that topological and geometrical measures can serve as potential markers for identifying Dengue outbreaks even in the context of high-level noise (in this case, associated with poor data acquisition or a constant endemic state). The sharpness of these markers varies according to the maximum cell dimension and time window size chosen.

Despite the results of our approach to synthetic data being satisfactory, the model chosen for generating and analysing data is a proof of concept only. Therefore, it is too simple to reproduce the realistic effects of Dengue behaviour and consider the noise generated from delays in the data collection. Yet a more comprehensive systematic analysis over a wider parameter range and different epidemic models deserves further investigation.

Even though, we believe that our results unveil a strong correlation between the Euler characteristic and the epidemic periods for both synthetic and real Dengue data. In fact, the presence of peaks (valleys) in the Euler characteristic preceding (or during) an epidemic peak may emerge as a warning sign for an outbreak. In analogy with the work in [43], where the authors used low dimensions of the Euler characteristic for detecting human brain activation areas from PET scan data, we suggest that the Euler characteristic close to Dengue epidemic periods might be used as markers of these events.

Ultimately, considering that topological data analysis is independent of the dynamics of the (epidemic) network, we believe our results will contribute to a better understanding of epidemic outbreaks from a topological point of view. In particular, our results can be further exploited and developed to improve next-generation alternative surveillance tools for epidemic outbreaks, which combine topological and geometrical approaches with other advanced data science techniques. One such possible further development of our approach in future studies could be to consider higher-order interactions (as opposed to pair-wise interactions) for generating simplicial complexes [87]. We believe that this approach might be an alternative topological surveillance method for vector-borne epidemic disease when the data set lacks robustness.

Acknowledgments

We thank Serafim Rodrigues at the Basque Center for Applied Mathematics and Katharina Natter at Leiden University for their critical review. D B S and F A N S proposed the research. D B S performed the research. F A N S supervised the research. D B S performed the data processing and the illustrations of the manuscript. D B S wrote the first version of the manuscript. E F S performed the interactive dashboards of the manuscript. All authors were involved in the subsequent discussions and development of the manuscript. D B S would like to also thank the Basque Government through the BERC 2018-2021 program and the Spanish State Research Agency through BCAM Severo Ochoa excellence accreditation SEV-2017-0718.

References

[1] Phan T 2020 *Novel coronavirus: From discovery to clinical diagnostics* Infection, Genetics and Evolution **79** 104211

[2] Lupia T, Scabini S, Pinna S M, Di Perri G, De Rosa F G and Corcione S 2020 *2019 novel coronavirus (2019-nCoV) outbreak: A new challenge* Journal of global antimicrobial resistance **21** 22–27

[3] Dias J P, Bastos C, Araújo E, Mascarenhas A V, Martins Netto E, Grassi F, Silva M, Tatto E, Mendonça J, Araújo R F *et al.* 2008 *Acute Chagas disease outbreak associated with oral transmission* Revista da Sociedade Brasileira de Medicina Tropical **41** 296–300

[4] Josseran L, Paquet C, Zehgnoun A, Caillere N, Le A T, Solet J L and Ledrans M 2006 *Chikungunya disease outbreak, Reunion Island* Emerging infectious diseases **12** 1994–1995

[5] Scallan E, Hoekstra R M, Angulo F J, Tauxe R V, Widdowson M A, Roy S L, Jones J L and Griffin P M 2011 *Foodborne illness acquired in the United States—major pathogens* Emerging infectious diseases **17** 7

[6] Naffakh N and Van Der Werf S 2009 *April 2009: an outbreak of swine-origin influenza A (H1N1) virus with evidence for human-to-human transmission* Microbes and infection **11** 725–728

[7] Bailey N T *et al.* 1975 *The mathematical theory of infectious diseases and its applications* (Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.)

[8] Brauer F 2005 *The Kermack–McKendrick epidemic model revisited* Mathematical Biosciences **198** 119–131

[9] Pastor-Satorras R, Castellano C, Van Mieghem P and Vespignani A 2015 *Epidemic processes in complex networks* Reviews of modern physics **87** 925

[10] Keeling M J and Eames K T 2005 *Networks and epidemic models* Journal of the Royal Society Interface **2** 295–307

[11] Newman M E 2002 *Spread of epidemic disease on networks*Physical review E **66** 016128

[12] Pastor-Satorras R and Vespignani A 2001 *Epidemic dynamics and endemic states in complex networks* Physical Review E **63** 066117

[13] Moore C and Newman M E 2000 *Epidemics and percolation in small-world networks* Physical Review E **61** 5678

[14] Jovanović M and Krstić M 2012 *Stochastically perturbed vector-borne disease models with direct transmission* Applied Mathematical Modelling **36** 5214–5228

[15] Britton T and Traoré A 2017 *A stochastic vector-borne epidemic model: Quasi-stationarity and extinction* Mathematical biosciences **289** 89–95

[16] Lord C, Woolhouse M, Heesterbeek J and Mellor P 1996 *Vector-borne diseases and the basic reproduction number: a case study of African horse sickness* Medical and veterinary entomology **10** 19–28

[17] Hartemink N, Cianci D and Reiter P 2015 *R0 for vector-borne diseases: Impact of the assumption for the duration of the extrinsic incubation period* Vector-borne and zoonotic diseases **15** 215–217

[18] Cao Y and Denu D 2016 *Analysis of stochastic vector-host epidemic model with direct transmission* Discrete & Continuous Dynamical Systems-B **21** 2109

[19] Zomorodian A 2012 *Topological data analysis* Advances in applied and computational topology **70** 1–39

[20] Wasserman L 2018 *Topological data analysis* Annual Review of Statistics and Its Application **5** 501–532

[21] Pascucci V, Tricoche X, Hagen H and Tierny J 2010 *Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications* Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications (Springer Science & Business Media)

[22] Taylor D, Klimm F, Harrington H A, Kramár M, Mischaikow K, Porter M A and Mucha P J 2015 *Topological data analysis of contagion maps for examining spreading processes on networks* Nature communications **6** 7723

[23] Edelsbrunner H and Harer J J 2010 *Computational topology: an introduction* (American Mathematical Society) ISBN 9780821849255

[24] Otter N, Porter M A, Tillmann U, Grindrod P and Harrington H A 2017 *A roadmap for the computation of persistent homology* EPJ Data Science **6** 17

[25] Bubenik P 2015 *Statistical topological data analysis using persistence landscapes* The Journal of Machine Learning Research **16** 77–102

[26] Saggar M, Sporns O, Gonzalez-Castillo J, Bandettini P A, Carlsson G, Glover G and Reiss A L 2018 *Towards a new approach to reveal dynamical organization of the brain using topological data analysis* Nature communications **9** 1–14

[27] Matamalas J T, Gómez S and Arenas A 2020 *Abrupt phase transition of epidemic spreading in simplicial complexes* Physical Review Research **2** 012049

[28] Amorim E, Moreira R A and Santos F A N 2022 *The Euler characteristic and topological phase transitions in complex systems* Journal of Physics: Complexity

[29] Sandhu R S, Georgiou T T and Tannenbaum A R 2016 *Ricci curvature: An economic indicator for market fragility and systemic risk* Science advances **2** e1501495

[30] Sandhu R, Georgiou T, Reznik E, Zhu L, Kolesov I, Senbabaoglu Y and Tannenbaum A 2015 *Graph curvature for differentiating cancer networks* Scientific reports **5** 12323

[31] Yen P T W and Cheong S A 2021 *Using topological data analysis (TDA) and persistent homology to analyze the stock markets in Singapore and Taiwan* Frontiers in Physics 20

[32] Santos F A, Raposo E P, Coutinho-Filho M D, Copelli M, Stam C J and Douw L 2019 *Topological phase transitions in functional brain networks* Physical Review E **100** 032414

[33] de Souza D B, da Cunha J T S, dos Santos E F, Correia J B, da Silva H P, de Lima Filho J L, Albuquerque J and Santos F A N 2021 *Using discrete Ricci curvatures to infer COVID-19 epidemic network fragility and systemic risk* Journal of Statistical Mechanics: Theory and Experiment **2021** 053501 URL `https://doi.org/10.1088/1742-5468/abed4e`

[34] Organization W H 2000 *Dengue: guidelines for diagnosis, treatment, prevention and control* World Health Organization **75** 193–196

[35] Villabona-Arenas C J and de Andrade Zanotto P M 2013 *Worldwide spread of dengue virus type 1* PloS one **8** e62649

[36] Hopp M J and Foley J A 2003 *Worldwide fluctuations in dengue fever cases related to climate variability* Climate Research **25** 85–94

[37] Cori A, Ferguson N M, Fraser C and Cauchemez S 2013 *A new framework and software to estimate time-varying reproduction numbers during epidemics* American journal of epidemiology **178** 1505–1512

[38] Thompson R, Stockwin J, van Gaalen R D, Polonsky J, Kamvar Z, Demarsh P, Dahlqwist E, Li S, Miguel E, Jombart T *et al.* 2019 *Improved inference of time-varying reproduction numbers during infectious disease outbreaks* Epidemics **29** 100356

[39] Organization W H, for Research S P, in Tropical Diseases T, of Control of Neglected Tropical Diseases W H O D, Epidemic W H O and Alert P 2009 *Dengue: guidelines for diagnosis, treatment, prevention and control* (World Health Organization)

[40] Bhatt S, Gething P W, Brady O J, Messina J P, Farlow A W, Moyes C L, Drake J M, Brownstein J S, Hoen A G, Sankoh O *et al.* 2013 *The global distribution and burden of dengue* Nature **496** 504–507

[41] Horstick O and Morrison A C 2014 *Dengue disease surveillance: improving data for dengue control* PLoS Neglected Tropical Diseases **8** e3311

[42] Wilder-Smith A and Byass P 2016 *The elusive global burden of dengue* The Lancet Infectious Diseases **16** 629–631

[43] Worsley K J, Evans A C, Marrett S and Neelin P 1992 *A three-dimensional statistical analysis for CBF activation studies in human brain* Journal of Cerebral Blood Flow & Metabolism **12** 900–918

[44] Ball F, Britton T, Leung K Y and Sirl D 2019 *A stochastic SIR network epidemic model with preventive dropping of edges* Journal of mathematical biology **78** 1875–1951

[45] Britton T, Juher D and Saldaña J 2016 *A network epidemic model with preventive rewiring: comparative analysis of the initial phase* Bulletin of mathematical biology **78** 2427–2454

[46] Benesty J, Chen J, Huang Y and Cohen I 2009 Pearson correlation coefficient *Pearson correlation coefficient* Noise reduction in speech processing (Springer) pp 1–4

[47] Fornito A, Zalesky A and Bullmore E T *Fundamentals of brain network analysis* ISBN 9780124081185

[48] Kim K, Kim S Y and Ha D H 2007 *Characteristics of networks in financial markets* Computer physics communications **177** 184–185

[49] Kenney J F and Keeping E S 1957 *Mathematics of statistics* Mathematics of statistics vol 2 (van Nostrand)

[50] Snedecor G W and Cochran W G 1937 *Statistical methods-80-7** Iowa State Press

[51] Edwards A L 1984 An introduction to linear regression and correlation Tech. rep.

[52] Erdos and P 1959 *On random graphs* Publicationes Mathematicae **6** 290–297

[53] Erdős P and Rényi A 1960 *On the evolution of random graphs* Publ. Math. Inst. Hung. Acad. Sci **5** 17–60

[54] Levy S 1999 *New perspectives in algebraic combinatorics* vol 38 (Cambridge University Press)

[55] Lott J and Villani C 2009 *Ricci curvature for metric-measure spaces via optimal transport* Annals of Mathematics 903–991

[56] Bochner S 1946 *Vector fields and Ricci curvature* Bulletin of the American Mathematical

Society **52** 776–797

[57] Colding T H 1997 *Ricci curvature and volume convergence* Annals of mathematics 477–501

[58] Tian G and Yau S T 1990 *Complete Kahler manifolds with zero Ricci curvature.* Journal of the American Mathematical Society **3** 579–609

[59] Matsumoto Y 2002 *An introduction to Morse theory* (American Mathematical Society) ISBN 9780821810224

[60] Bobenko A 2008 *Discrete differential geometry* (Springer)

[61] Forman R 2003 *Bochner's method for cell complexes and combinatorial Ricci curvature* Discrete and Computational Geometry **29** 323–374

[62] Sreejith R, Mohanraj K, Jost J, Saucan E and Samal A 2016 *Forman curvature for complex networks* Journal of Statistical Mechanics: Theory and Experiment **2016** 063206

[63] Ollivier Y 2009 *Ricci curvature of Markov chains on metric spaces* Journal of Functional Analysis **256** 810–864

[64] Pouryahya M, Mathews J and Tannenbaum A 2017 *arXiv preprint arXiv:1712.02943*

[65] Knill O 2011 *A graph theoretical Gauss-Bonnet-Chern theorem* arXiv preprint arXiv:1111.5395

[66] Lundell A T and Weingram S 2012 *The topology of CW complexes* (Springer Science & Business Media)

[67] Carlsson G 2009 *Topology and data* Bulletin of the American Mathematical Society **46** 255–308 ISSN 0273-0979

[68] Amorim Filho E C d 2019 *Topological transitions on protein-protein interaction networks* Universidade Federal de Pernambuco

[69] Hatcher A 2005 *Algebraic topology* (Cambridge University Press)

[70] Greenberg M J 2018 *Algebraic topology: a first course* (CRC Press)

[71] Knill O 2011 *On the Dimension and Euler characteristic of random graphs* URL `https://arxiv.org/pdf/1112.5749.pdf`

[72] Heiss T and Wagner H 2017 *Streaming algorithm for euler characteristic curves of multidimensional images* International Conference on Computer Analysis of Images and Patterns (Springer) pp 397–409

[73] Smith A and Zavala V M 2021 *The Euler characteristic: A general topological descriptor for complex data* Computers & Chemical Engineering **154** 107463

[74] Knill O 2012 *On index expectation and curvature for networks* arXiv preprint arXiv:1202.4514

[75] Knill O 2012 *An index formula for simple graphs* arXiv preprint arXiv:1205.0306

[76] Knill O 2014 *Curvature from graph colorings* arXiv preprint arXiv:1410.1217

[77] Knill O 2013 *The Euler characteristic of an even-dimensional graph* arXiv preprint arXiv:1307.3809

[78] Stolerman L M, Coombs D and Boatto S 2015 *Sir-network model and its application to dengue fever* SIAM Journal on Applied Mathematics **75** 2581–2609

[79] Baez J C and Biamonte J 2018 *Quantum techniques in stochastic mechanics* (World Scientific)

[80] Erban R, Chapman J and Maini P 2007 *A practical guide to stochastic simulations of reaction-diffusion processes* arXiv preprint arXiv:0704.1908

[81] `https://www.python.org`

[82] Annunziato A and Asikainen T 2020 *Effective reproduction number estimation from data series* JRC121343

[83] `https://www.kaggle.com/datasets/danillosouza2020/synthetic-epidemic-vs-rt-and-euler-cha`

[84] `https://kaggle.com/datasets/danillosouza2020/dengue-data-and-its-euler-characteristics`

[85] `http://dados.recife.pe.gov.br/dataset/casos-de-dengue-zika-e-chikungunya`

[86] `https://www.kaggle.com/datasets/danillosouza2020/recife-denague-data-vs-rt-and-euler-ch`

[87] Battiston F, Cencetti G, Iacopini I, Latora V, Lucas M, Patania A, Young J G and Petri G 2020 *Networks beyond pairwise interactions: structure and dynamics* Physics Reports **874** 1–92