IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 23, 2021, accepted May 17, 2021, date of publication May 19, 2021, date of current version May 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3082014

Assessing Granger Causality on Irregular Missing and Extreme Data

MASSIMILIANO ZANIN
Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Campus UIB, 07122 Palma de Mallorca, Spain
e-mail: massimiliano.zanin@gmail.com

This work was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme under Grant 851255, and in part by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia e Innovación (MCI), Spain, and the Fondo Europeo de Desarrollo Regional (FEDER, UE), under the María de Maeztu Program for units of Excellence in Research and Development (MDM-2017-0711).

ABSTRACT The Granger test is one of the best known techniques to detect causality relationships among time series, and has been used uncountable times in science and engineering. The quality of its results strongly depends on the quality of the underlying data, and different approaches have been proposed to reduce the impact of, for instance, observational noise or irregular sampling. Less attention has nevertheless been devoted to situations in which the analysed time series are irregularly polluted with missing and extreme values. In this contribution I tackle this problem by comparing four different data pre-processing strategies and evaluating their performance with synthetic time series, both in dyadic tests and functional network contexts. I further apply these strategies to a real-world problem, involving inferring the structure behind the propagation of delays in an air transport system. Finally, some guidelines are provided on when and how these strategies ought to be used.

INDEX TERMS Granger causality, missing data, functional networks.

I. INTRODUCTION

The Granger causality test [1], developed by the economy Nobel Prize laureate Clive Granger on top of time series prediction models proposed by Norbert Wiener [2], is one of the best well-known metrics for assessing *predictive causality* [3] between elements composing a system. It is based on the analysis of time series, and on a very simple and intuitive assumption: given two elements A and B, B is causing A if including information about the past of B helps predict the future of A - as, in other words, B contributes in defining the future of A. Since its introduction, this test has been applied to uncountable problems, from economics [4]–[7], engineering [8], sociology [9], biology [10] or neuroscience [11]–[13]; and has been extended to handle different situations and types of data [14]–[17].

As is ubiquitous in data analysis, results obtained by the Granger causality test are as good as the data supporting them. Any practitioner working on real-world problems knows that data are seldom (if ever) perfect; but that instead they usually contain different types of artefacts. To illustrate, and following the previous examples, financial markets can close because of local public holidays; when merging data from different countries, this may result in gaps in the time series. Similarly, any application of Granger causality in engineering is based on recording variables from sensors, which are subject to measurement and additive noise and that may be sampled at a frequency not corresponding with the natural one of the system. Finally, biological data cannot always be recorded, resulting in irregularly-sampled time series.

The challenges posed by these data limitations in the application of the Granger causality test have been the focus of previous works in the literature. Specifically, researchers have proposed solutions to improve the detection of causality under noise [18]–[20], linear transformations and subsampling [21], irregularly sampled time series [22], and equidistant missing data [23].

Less attention has been devoted to the problem of how Granger causality behaves under (irregular) missing and extreme values. Note that irregular pollution of data is a scenario more common that, for instance, a regular one, as studied in [23]. For the sake of simplicity, here missing values are defined as those measurements that are not available and are thus encoded by zeros; and extreme values as those

The associate editor coordinating the review of this manuscript and approving it for publication was Haluk Eren.

that fall beyond the expected range of normal measurements. Several general methods, i.e. not specific for Granger causality, have been proposed in the past to handle such values. These include, for instance, deletion, i.e. simply deleting those instances containing one or more missing data [24]; interpolation, in which missing values are supposed to be a function of the neighbouring (non-missing) ones [25]; or random replacement, in which missing values are filled with random numbers drawn from the original time series [26]. These methods, while useful in general, may not be so when applied to Granger causality. The reason stems from the fact that this test goes beyond considering values of a time series as independent from each other, but instead assesses temporal structures. To illustrate, substituting missing values with data drawn from the time series guarantees that the probability distribution is maintained, and may be suitable in many applications [27]. This method nevertheless yields suboptimal results when used in conjunction with Granger causality, as I will show below, as preserving the probability distribution is not enough to preserve temporal relationships. In addition, there is an increasing interest in the scientific community in evaluating large-scale causality structures, i.e. beyond simple dyadic relationships [28]. In other words, given a set of variables, the Granger test is calculated between each pair of variables, and the final structure is then described in terms of a set of network metrics. This is especially relevant in cases where multiple elements are expected to interact in complex ways, as is for instance the case of brain regions [11], [12]. How these standard methods affect the observed network structures is something that has not yet been studied.

In this contribution I review and evaluate four methods for handling missing and extreme values in conjunction with Granger causality: doing nothing, i.e. disregarding the presence of wrong values; substituting those values with the median of the time series, also known as a value imputation strategy; substituting them with random ones, known as random replacement; and a novel method based on weighting the linear regression model underlying the Granger causality test, in order to disregard missing and extreme data. These methods are tested using synthetic data, to detect both how many real causal relations they are able to recover and how many spurious relations are generated. This is done both in a bivariate (or dyadic) case, i.e. when pairs of time series are independently tested; and in a functional network context [29], in which the Granger causality test is used to detect connections between the nodes composing a network [30], [31]. The results of these tests on synthetic data are used to draw some guidelines, and are further applied to a real-world application involving the reconstruction of causal networks representing the propagation of delays in the European air transport system [32].

II. THE GRANGER CAUSALITY TEST AND ITS STABILITY

While the Granger causality test is a well-known instrument used in many fields of science, and has been described in uncountable publications, for the sake of completeness the main elements of its mathematical formulation are discussed here below. Suppose two systems A and B, respectively described by two time series t_A and t_B representing some observable function of their dynamics. These two systems are part of the universe U, representing all systems and elements (both observable by and hidden to the researcher) that are relevant for a given problem. Let us further suppose that these time series fulfil some basic conditions, including being stationary and regularly sampled. Note that the F- and χ^2-tests underlying the Granger approach assume stationarity; therefore, if a unit root is observed in the time series, it is recommended to make them stationary by using the first- or second-order time derivative, as otherwise spurious causalities may emerge. Additionally, for a discussion on irregularly sampled time series, see [22]. B is said to "Granger-causes" A if:

$$\sigma^2(t_A|U^-) < \sigma^2(t_A|U^-\backslash t_B^-), \tag{1}$$

where $\sigma^2(t_A|U^-)$ stands for the error, in terms of the standard deviation of residuals, when forecasting the time series t_A using the past information of the entire universe U; and $\sigma^2(t_A|U^-\backslash t_B^-)$ the error when the information about time series t_B is removed. In other words, B is causing A if including information about the past of B helps predict the future of A - hence the name of *predictive causality* [3]. The forecast itself is done, in both cases, through an autoregressive-moving-average (ARMA) model. An F-test is finally performed to assess the statistical significance of the inequality of Eq. 1; a causality is accepted if the resulting p-value is below a fixed significance value (usually $\alpha = 0.01$). Note that other alternatives to the F-test are available, as for instance the Wald test, which assesses the absence of causality (or, equivalently, assesses the noncausality) [33]; results here presented, unless otherwise specified, correspond to the F-test.

How stable is the Granger causality test? In other words, how resilient are its results when missing and extreme values are included? In order to answer this initial question, I here consider a simple linear system, composed of two time series:

$$\begin{cases} x(t) = \xi_x \\ y(t) = \gamma x(t-2) + \xi_y, \end{cases} \tag{2}$$

with ξ_x and ξ_y representing two independent and identically distributed random variables with probability distribution $\mathcal{N}(0, 1)$, and 2 the *lag*, i.e. the time required for x to force y. Note that x has an independent dynamics, while y dynamics is partly defined by the past of x through a coupling constant γ; thus, provided γ is large enough, the Granger test should be able to detect such causality relation. As a final step, a fraction δ of the values of both time series is randomly selected and set to zero, to simulate the presence of missing values. Note that missing values are here distributed in an irregular way - for a discussion of equidistant missing data, see [23].

Fig. 1 reports the fraction of detected causality relations in these raw time series as a function of δ and γ; in other words, it reports the fraction of times the Granger test detected

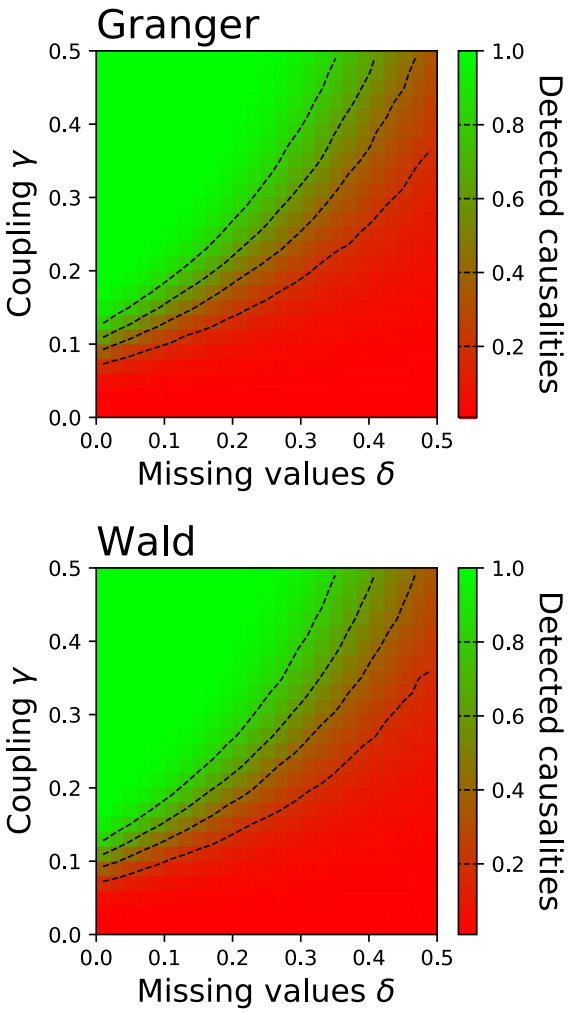

FIGURE 1. Resilience of the Granger causality test to missing values. Evolution of the fraction of detected causalities (i.e. fraction of times the Granger causality yielded a statistically significant result) as a function of the fraction of missing values δ and of the coupling γ, for time series defined as per Eq. 2. The top and bottom panels respectively correspond to a standard Granger test, and to the Wald test variant. For each pair of values (δ, γ), 10^4 simulations have been executed, and the fraction of simulations yielding a p-value below $\alpha = 0.01$ is here reported.

a statistically significant causality. The top and bottom panels respectively correspond to a standard Granger test (thus based on an F-test) and a Wald test variant. For a coupling strength $\gamma \approx 0.1$, the Granger test is able to detect the presence of a causality relation half of the times, provided no values are missing; yet, for the same coupling strength, almost no tests yield statistically significant results for $\delta > 0.1$. Additionally, virtually the same results are obtained when using the Wald test [33], suggesting that missing values have the same effect independently on the used test.

Beyond this simple bivariate example, as previously introduced, the Granger causality test is frequently used to reconstruct functional networks, i.e. graphs representing the structure of interactions between the elements of a complex system [29]. In order to explore this case, the system of Eq. 2 has been extended to include $N = 40$ elements (or nodes), with the dynamics of element i-th being defined by:

$$x_i(t) = \xi + \gamma \sum_{j=1}^{N} a_{j,i} x_j(t-2). \tag{3}$$

The matrix $\mathcal{A}$ is commonly called the *adjacency matrix*, and its element $a_{j,i}$ is equal to one whenever node j is connected to (here, forces) node i, and zero otherwise [30], [31]. The dynamics of each element is thus the result of the sum of an internal component, in this case stochastic; and of external contributions defined by the elements of $\mathcal{A}$. I here consider that the probability of connecting two nodes i and j is proportional to $1/i$, such that node 1 is the most connected one, and node 40 is the least one.

In a real-world application, what is available to the researcher is the set of time series x_i, and the aim is to reconstruct the underlying functional network and its adjacency matrix $\mathcal{A}$ by applying the Granger causality test on each pair of time series. Note that, while being the most common approach, it is not the only one; for instance, the Granger causality can be estimated through a vector autoregression (VAR) model to model multivariate (i.e. beyond dyadic) dependencies [34]. In both cases, the interest is shifted from evaluating the presence of individual relationships, to the assessment of high-level structures emerging from such micro-scale - what is commonly called the *topology* of the network [35]. As such, I here consider four classical topological metrics used in network science:

- *Link density.* Number of active links divided by the total number of possible links - in a directed network of N nodes, this latter number is $N(N-1)$:

$$l_d = \frac{1}{N(N-1)} \sum_{i,j} a_{i,j} \tag{4}$$

- *Transitivity.* Density of triangles in the network, i.e. of triplets of nodes such that, if A is connected to B and this to C, then C is also connected to A [36]. Also called *clustering coefficient*, this metric is defined in terms of the adjacency matrix as:

$$C = \frac{\sum_{i,j,k} a_{i,j} a_{j,k} a_{k,i}}{\sum_i k_i (k_i - 1)}, \tag{5}$$

  where k_i is the number of connections (i.e. the *degree*) of node i.
- *Efficiency.* Metric describing how efficiently the network exchanges information, and defined as the inverse of the harmonic mean of the minimum distance between all pairs of nodes [37].
- *Assortativity.* Correlation coefficient between the degrees of nodes at the extreme end of each link [38]. Positive values indicate that the network is *assortative*, and that highly connected nodes tend to connect with themselves. On the other hands, networks with negative values of the metric are called *disassortative*, and are

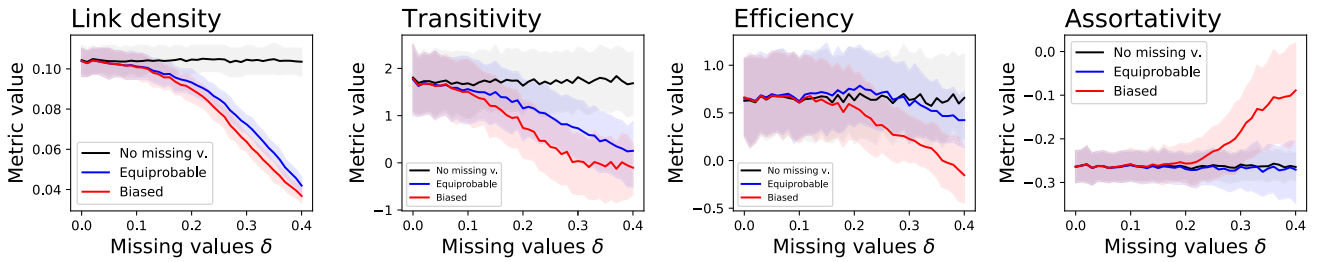

FIGURE 2. Evolution of four topological metrics, for networks reconstructed from time series defined as per Eq. 3, as a function of the fraction of missing values δ and for $\gamma = 0.3$. Black, blue and red lines respectively correspond to no missing values; missing values equally distributed among time series; and missing values biased towards some time series. Deviations from the black lines thus indicate a bias in the observed metric values. Each point corresponds to the average of 500 random realisations; transparent bands to the $25-75$ percentile band.

characterised by having highly connected nodes being attached preferentially to peripheral ones.

It is important to note that both the transitivity and the efficiency depend on the number of links in the network. To illustrate, a dense network will have a higher efficiency that a sparse one, as a larger number of links implies a higher probability of having a direct path between pairs of nodes. Yet, this does not imply that links are organised in a more efficient way, or in a way favouring movements in the network. In order to compare networks with different number of links, the two metrics have to be transformed through a Z-Score, i.e. the distance between the observed metric and what expected in an ensemble of random equivalent (same number of nodes and links) graphs [39]. In what follows, reported values of transitivity and efficiency correspond to their respective Z-Scores.

Fig. 2 presents the evolution of the average value of these four topological metrics, as obtained from reconstructing the functional network defined in Eq. 3, as a function of δ for three different cases. First of all, the blue lines correspond to a case in which missing values are distributed among the time series in an equiprobable way. On the other hand, this probability is made proportional to i (i.e. to the node number) in the case of the red lines; in other words, the last nodes of the set, which are also the least connected ones, receive the largest share of missing values. Finally, the black lines correspond to the network reconstructed on the data without missing values, and thus represent the ideal result. Any deviation from the black lines indicates a bias in the observed topological metrics.

It can be appreciated that results strongly vary, and depend on the considered topological metric. For instance, missing values have no effect on the efficiency and assortativity of the network, provided these are few and distributed equally among the time series. As long as missing values are not biased, these are causing random links to disappear, thus they do not affect (or bias) the overall structure of the network. This changes both when a large fraction of values (larger than approximatively 20%) are missing, as the network structure is effectively destroyed; and when missing values are not equally distributed, as this introduces a bias in the results.

In synthesis, the results of Figs. 1 and 2 describe a complex picture. On one hand, the Granger causality test is quite sensitive to missing values, as even a small fraction of them can undermine the detection of weak and medium causalities. When the focus is shifted towards large-scale structures, some topological metrics are rather insensitive to missing values, provided these are few and evenly distributed among time series.

III. HANDLING MISSING AND EXTREME VALUES DATA

A. STRATEGIES DEFINITION

In this contribution I am considering three strategies for handling extreme and missing values, plus an initial baseline reference. These are:

- *Using Raw Time Series:* The baseline of the analysis is obtained by disregarding that some values are missing or extreme, and thus work with the raw time series. As previously shown, this approach yields a substantial subestimation of the causality in bivariate analyses, and changes in topological metrics for functional networks.
- *Substituting Missing and Extreme Values With the Median:* This second strategy is based on substituting wrong values for values that are expected to less mislead the calculation of the Granger causality. Specifically, I here firstly consider the median of all available values, the rationale being that this should affect less the calculation of the slope coefficients in the linear regression model. In other words, encountering a (false) value equal to the median pushes the slope towards zero, but does not introduce any bias. This strategy is thus a value imputation one.
- *Substituting Missing and Extreme Values With Random Ones:* This strategy is equivalent to the previous one, except that wrong values are substituted by values drawn from the same time series. This is a random replacement strategy that has widely been used in other contexts, and which presents the advantage of maintaining the underlying probability distribution [26].
- *Weighting the Linear Regression Model:* Instead of fitting the linear model of the Granger causality considering equal weights for all values, one may use

a linear model in which the weight of missing and extreme values is set to zero. In other words, consider the simplest case in which the future of a time series X is forecasted through an autoregressive model of its past, i.e. $x_t = a_0 + a_1\, x_{t-1} + a_2\, x_{t-2} + \ldots + \epsilon_t$, with ϵ_t representing the error of the forecast. The coefficients a can be obtained through an ordinary least squares approach. Subsequently, let us suppose that a value x_j is known to be wrong, e.g. missing or extreme. The previous fit can then be substitute by a weighted least squares approximation, in which all weights w are set to one, except for w_j that is set to zero. This effectively implies that the value of x_j is excluded from the calculation of the autoregressive model. When this approach is applied to the two models composing the two sides of Eq. 1, the result is a Granger causality test performed only on the values esteemed as correct. Note that this approach also presents the drawback of a decrease in the reliability of subsequent F-tests, as the effective length of the time series is changed. To the best of my knowledge, this method has not been studied before.

Note that some other common strategies are here not considered, as they depend on the characteristics of the data, and are therefore not always relevant. These include, on one hand, interpolating missing and extreme values using neighbouring ones [25]. This strategy clearly assumes that values are not independent, or, in other words, that the time series have a non-zero autocorrelation; nevertheless, this also means that time series are not stationary, and therefore that the Granger causality test may be overestimating the results. On the other hand, one may consider trimming the time series, in order to isolate "clean" parts of them, or even deleting those time series containing missing values; these two approaches respectively require to have few missing values (or for them to be clustered in a single region), and to have multiple instances of each time series.

In order to see how the four described strategies can help solving the missing and extreme values problem, I here start by applying them to a simple correlation. Even though the Granger causality test is more than a simple linear correlation, the latter is a basic element of the former, at least when an ARMA model is used in its estimation. For this I consider a simple linear system, composed of two time series $x = [1, 2, \ldots, 20]$ and $y = x + \xi$, where ξ represents independent random numbers drawn from a normal distribution $\mathcal{N}(0, 1)$. As in the case of Eqs. 2 and 3, a fraction δ of the values is randomly selected and set to zero, to simulate the presence of missing values. Finally, a linear regression model, and the corresponding p-value of the F-test to check whether the slope is significantly different from zero, are calculated. Fig. 3 reports the median of the correlation coefficient (top panel) and of the $\log_{10}$ of the p-value (bottom panel) over $10,000$ executions. It can be appreciated that, as expected, using the raw time series and disregarding the fact that some values are missing underestimates the correlation coefficient, and the corresponding p-value converges to 1 for large values

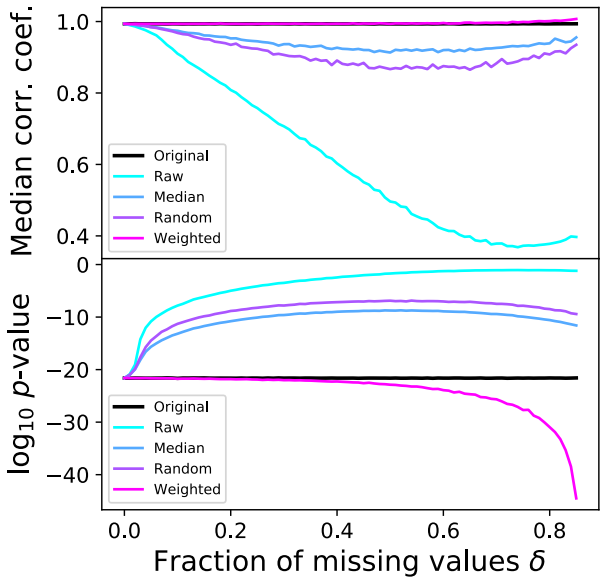

FIGURE 3. Evolution of the median of the linear correlation coefficient (top panel) and of the corresponding p-value (bottom panel), as a function of the fraction of missing values δ, and for the four strategies here consider (see legend for colour codes). Results correspond to 10^5 random realisations. See main text for a definition of how time series are generated.

of δ (aqua line). The best method seems to be the weighted linear model, which yields a correlation coefficient very close to that of the original time series without missing values (black line); nevertheless, starting from $\delta \approx 0.5$, the p-value is strongly underestimated. This result is easy to be visualised, by imagining an extreme situation in which only two values are not missing from both time series: the correlation is then calculated over these two pairs of values, yielding a perfect fit - and hence a subestimation of the true p-value. Finally, the two remaining methods based on substitutions perform similarly, retaining a good and statistically significant approximation of the correlation coefficient.

B. TESTING PAIR-WISE PREPROCESSING STRATEGIES

I then move to analyse how these strategies perform in a real causality detection problem, focusing in a bivariate (or dyadic) case involving the detection of the causality between a pair of time series. For that, I will use again the model of Eq. 2. Unless otherwise specified, I here consider time series of $1,000$ points and $\gamma = 0.1$, yielding $\approx 60\%$ of statistically significant tests when no wrong values are included - for a significance level of $\alpha = 0.01$. Note that the value of $\gamma = 0.1$ has been selected to represent an intermediate scenario, in which causality relationships are neither too strong, as this would minimise the impact of missing values, nor too weak, as to preclude a detection.

Fig. 4 depicts the result of the Granger causality test as a function of the fraction δ of missing values and of the data preprocessing method. In a way similar to Fig. 3, two set of

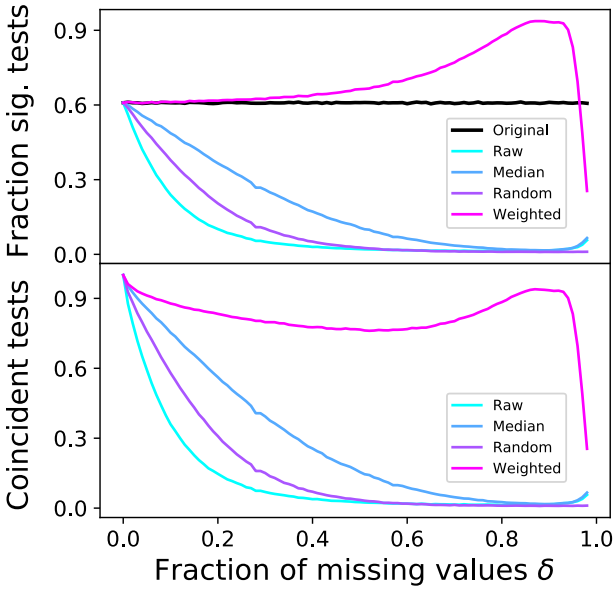

FIGURE 4. Effect of missing data handling strategies on bivariate Granger causality tests. The top panel depicts the fraction of tests identified as significant, for each one of the four strategies, and as a function of the fraction of missing values δ. The bottom panel further reports, for each strategy, the fraction of tests that were also significant in the original (i.e. without missing values) time series. Results correspond to $\gamma = 0.1$, time series of length 1,000 and 10^5 random realisations.

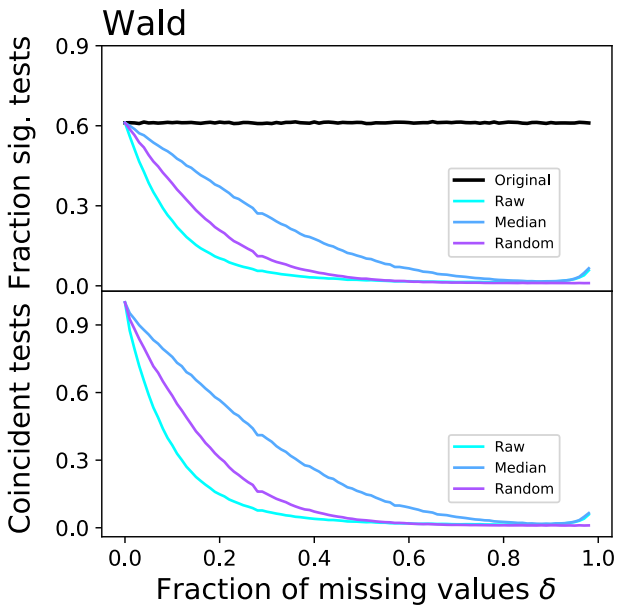

FIGURE 5. Effect of missing data handling strategies on bivariate Granger causality tests, when calculated using the Wald test - see main text, Sec. II for details. Meaning of panels and colours is as per Fig. 4.

results are presented: the fraction of statistically significant tests (top panel), and the fraction of time each pre-processing method yields the same result as the one corresponding to the original time series (bottom panel). Note that the result of the Granger causality on the original (i.e. without missing values) time series is here considered as the *gold truth*, i.e. the result that one would like to recover in spite of the missing values. Also note that this is similar but not equivalent to what presented in Fig. 3, as here the statistical significance of the test refers to a joint linear hypothesis in which at least one correlation coefficient, i.e. for at least one time lag, is different from zero.

Several interesting results can be observed. First of all, and confirming what observed in Fig. 1, disregarding the fact that some values are missing results in a dramatic underestimation of the causality - aqua line in Fig. 4. On the other hand, of the methods here studied, weighting the regression model is the best performing one, especially for $\delta < 0.4$; still, and in line with what observed in Fig. 3, it yields an overestimation of the causality when the majority of values are missing. Finally, of the two remaining methods, the use of the median is the better option, still detecting half of the significant tests for δ as large as 0.3.

The generality of these results are next evaluated using three additional models. Firstly, Fig. 5 presents the results corresponding to the same coupled linear model of Eq. 2, but here using the Wald test [33] - see also Sec. II for a definition. Secondly, Fig. 6 Left depicts the causality detected in a biologically inspired model, specifically synthetic time series emulating the signal recorded by function Magnetic Resonance Imaging (fMRI) in the human brain. Two coupled time series are initially generated through a Vector AutoRegressive (VAR) stochastic model and additive Gaussian noise; these are then convoluted with the canonical haemodynamic response function (HRF), defined as a mixture of two gamma functions; the results are finally downsampled - for a full definition, please refer to [40], [41]. Thirdly, Fig. 6 Right corresponds to time series generated by two coupled Lorenz oscillators, a prototypical example of chaotic oscillator [42], and defined respectively as:

$$\dot{x_1} = \sigma(y_1 - x_1) \tag{6}$$
$$\dot{y_1} = x_1(\rho_1 - z_1) - y_1 \tag{7}$$
$$\dot{z_1} = x_1 y_1 - \beta z_1 \tag{8}$$

and

$$\dot{x_2} = \sigma(y_2 - x_2) - \alpha(x_2 - x_1) \tag{9}$$
$$\dot{y_2} = x_2(\rho_2 - z_2) - y_2 - \alpha(y_2 - y_1) \tag{10}$$
$$\dot{z_2} = x_2 y_2 - \beta z_2 - \alpha(z_2 - z_1), \tag{11}$$

with $\sigma = 10.0$, $\rho_1 = 28$, $\rho_2 = 29$, and $\beta = 2.667$. Note that oscillator number 2 is driven by oscillator number 1 through the coupling constant α (here set to 0.5). The two time series here considered correspond to the channels x_1 and x_2. In all these three cases, the considered time series have a length of 1,000 points.

Both set of results are qualitatively similar to Fig. 4, and even almost undistinguishable in the case of Fig. 5. It is nevertheless worth noting the drop in the fraction of significant tests at $\delta \approx 0.2$ in the case of fMRI synthetic signals, and at $\delta < 0.05$ in the case of the Lorenz oscillators.

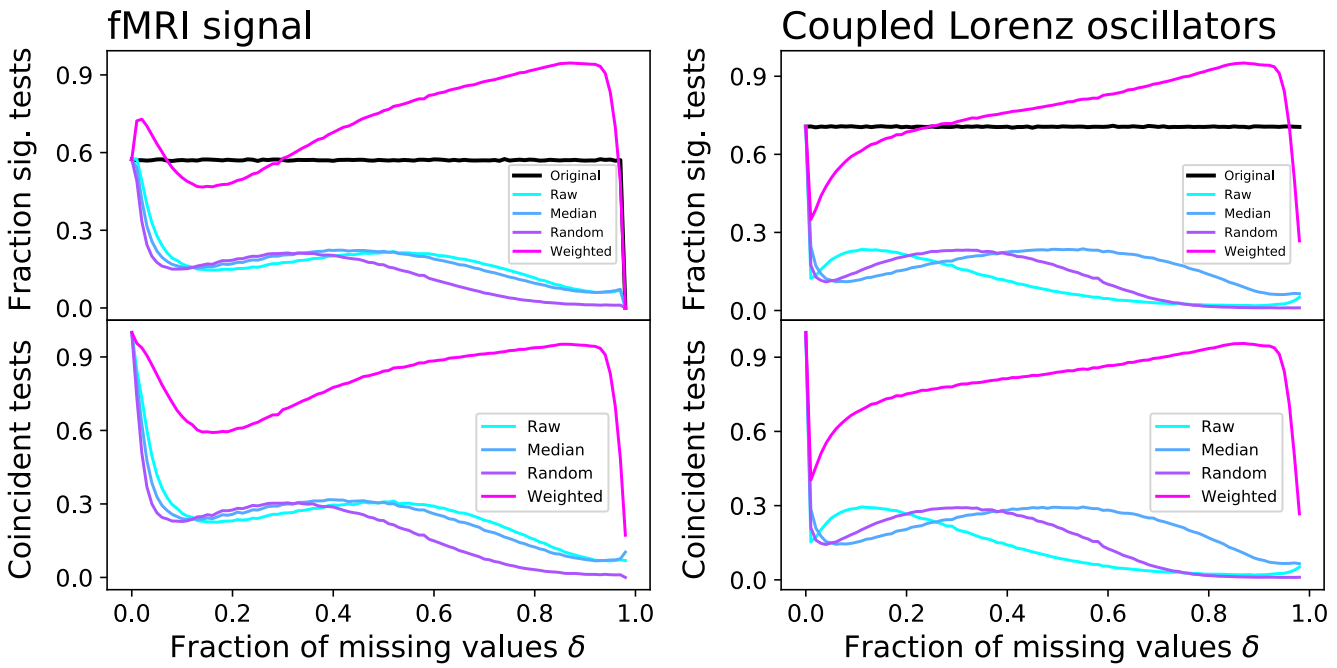

FIGURE 6. Effect of missing data handling strategies on bivariate Granger causality tests, calculated over synthetic time series representing the dynamics of coupled brain regions as measured by function Magnetic Resonance Imaging (fMRI, left panel), and coupled Lorenz oscillators (right panel). Meaning of panels and colours is as per Fig. 4.

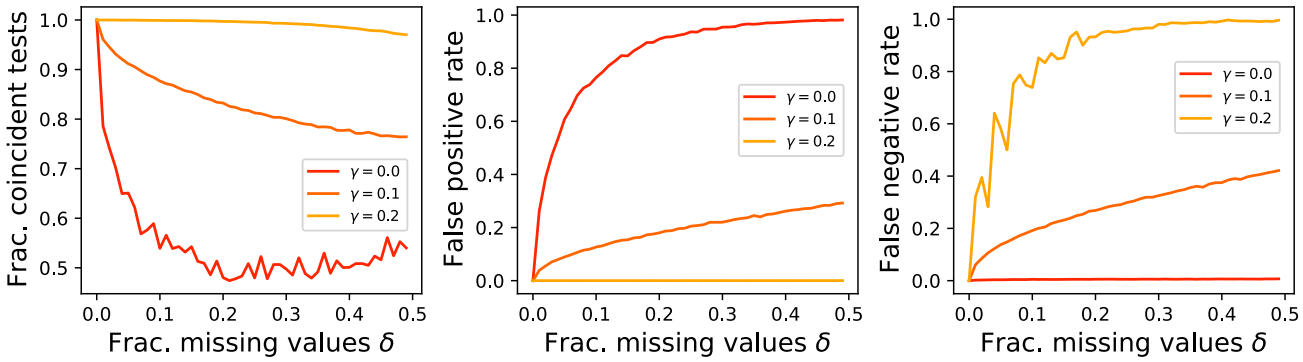

FIGURE 7. Evaluation of the model weighting strategy. The three panels report, from left to right and as a function of the fraction of missing values δ: the fraction of correct results, the false positive rate, and the false negative rate. Red, orange and yellow lines respectively correspond to $\gamma = 0$ (i.e. no coupling between the time series), $\gamma = 0.1$ and $\gamma = 0.2$.

In order to understand the overestimation seen for the weight method in Fig. 4, Fig. 7 presents the evolution as a function of γ and δ of three additional metrics: (i) the fraction of times the result coincide with the analysis of the original time series (left panel); (ii) the false positive rate (center panel); and (iii) the false negative rate (right panel). Let us first consider the case of $\gamma = 0$, i.e. when the time series x and y are completely independent; in this case the number of significant tests should be approximatively equal to the significance level α, here 0.01. For large values of δ, there are many false positive cases, and very few false negatives; in other words, the weighting method overestimates the presence of a causality relation (which is actually not present at all). Most notably, the opposite happens for larger values of γ, as for instance for $\gamma = 0.2$: the weighted test almost perfectly recovers the results for the original time series, yielding a larger number of false negatives; in other words, it errs towards an underestimation of the causality. This behaviour is stronger for even larger values of γ - not shown here for the sake of clarity.

If the previous results corresponded to synthetic data sets with missing data, one may wonder if similar outcomes are achieved in the case of extreme values. For that, one can again start from time series created through the model of Eq. 2; for then selecting a fraction δ of elements at random, and substituting them with random numbers drawn from

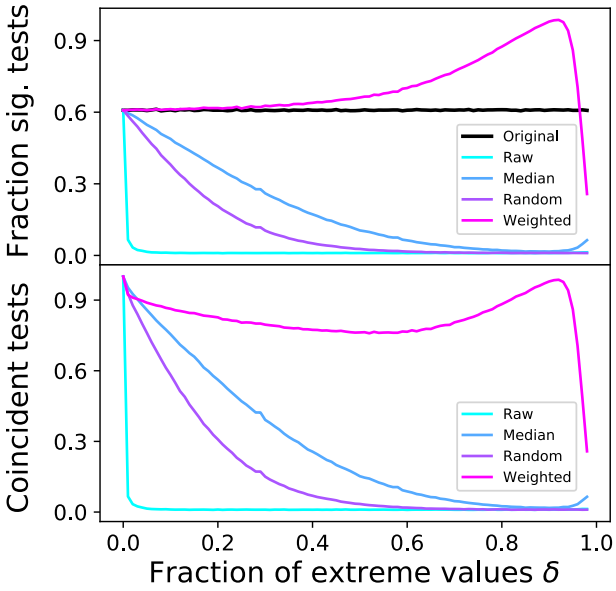

FIGURE 8. Effect of extreme values handling strategies on bivariate Granger causality tests. Top and bottom panels respectively depict the fraction of tests identified as significant, and the fraction of tests that were also significant in the original (i.e. without extreme values) time series. Colour code as per Fig. 4. Results correspond to $\gamma = 0.1$ and 10^5 random realisations.

a distribution $\mathcal{N}(10, 1)$. Note that, as stated in the introduction, missing and extreme values are here defined as those values that are wrong, with the former ones being encoded by values *within* the distribution (usually, zeros), and the latter ones by values *outside* it. Fig. 8 reports the results for this modified model, using the same schema and colour code as Fig. 4. It can be appreciated that results are qualitatively the same, with the weighted model being the best strategy to recover lost causality relationships - albeit at a cost of an overestimation for large values of δ. From the point of view of handling wrong data in Granger causality tests, missing and extreme values seem to be equivalent.

C. TESTING PREPROCESSING STRATEGIES, NETWORK EFFECTS

Once the four pre-processing strategies have been tested on bivariate time series, the next natural step is to ascertain whether they are able to recover the lost values of topological metrics. Towards this aim, these strategies have been applied to the system defined in Eq. 3, with $\gamma = 0.3$ and a biased distribution of missing values, thus according to the red lines of Fig. 2. The results are presented in Fig. 9; the same colour code as in Fig. 2 has been used, such that deviations from the black lines indicate a bias in the results. These results point to a complex picture. Firstly, the approach based on weighting the linear model is able to compensate for the changes in structure introduced by missing values, even though the transitivity seems to be underestimated for $\delta > 0.1$.

Secondly, substituting missing values for the median is not introducing a significant improvement (note that its lines, dark blue, almost perfectly coincide with the ones of the raw data, light blue). While this seems surprising at first, it is easy to identify the reason behind such result. Specifically, Eq. 3 indicates that each time series is composed of values normally distributed around zero (plus coupling signals, which also have zero mean). Therefore, when missing values (encoded as zeros) are substituted by the median of the distribution (approximatively zero), the net effect is negligible. This highlights that the performance of this method (and, eventually, of all methods) depend not just on the density of missing values, but also on the probability distribution of the underlying data, a topic that will be further discussed in the conclusions.

Finally, using a random replacement strategy increases the biases in the observed structures, with respect to using the raw (non pre-processed) data. Therefore, this strategy ought to be avoided in a network context.

D. CONCLUSION: IS THERE AN OPTIMAL STRATEGY?

Combining all the results obtained on these simple synthetic data sets allows to draw some conclusions, on how missing and extreme data should be handled when assessing causalities through the Granger test.

- The existence of missing and extreme values cannot be disregarded, unless they are few and the expected strength of the causal connection is large. As seen in Fig. 1, a mere 10% of missing values can have catastrophic effects if the causality relation is weak. When the focus is shifted towards functional networks and their topological metrics, the latter ones are somewhat more resilient, but are also affected by the way missing values are distributed among the time series.
- The Granger causality test behaves similarly under the presence of missing and extreme values, i.e. wrong measurements respectively falling inside and outside the expected data probability distribution - see Fig. 8.
- If missing and extreme values represent more than 20% of all available data, the only real solution is to resort to a weighted model for calculating the Granger causality. This comes at the cost of overestimating weak causalities, and underestimating strong ones. Still, around 80% of the significant dyadic relations can be recovered (see Fig. 4 right panel). Network metrics are generally, but not always, correctly recovered - see e.g. the underestimation of the transitivity in Fig. 9. In any case, results obtained from time series with more than 20% of errors have to be interpreted with due care.
- For fractions of missing values below 20%, the best option is still the weighted model. Nevertheless, in cases requiring a conservative estimation (that is, when the cost of obtaining false positives is much larger than the one of false negatives) and for dyadic relationships, an alternative solution is to substitute missing values with the median, which allows to recover between 60 and 80% of the causality links.
- Finally, a random replacement strategy is generally to be avoided.

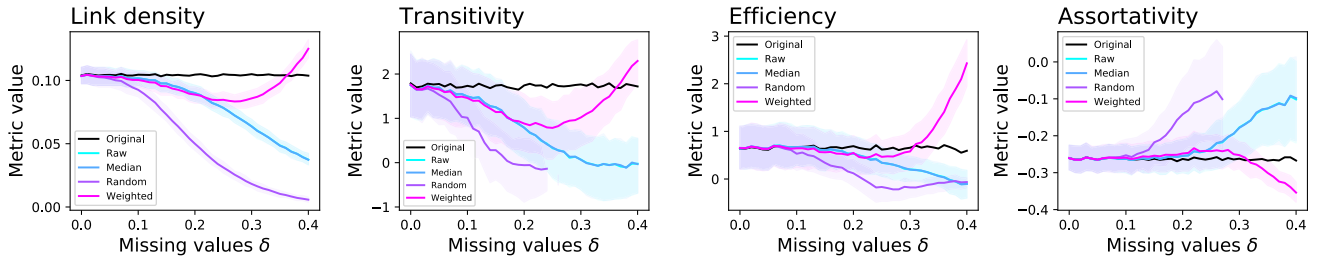

FIGURE 9. Evolution of four topological metrics, as obtained after applying the four strategies here considered, as a function of the fraction of missing values δ and for $\gamma = 0.3$. As in Fig. 2, deviations from the black lines (representing the result without missing values) indicate biases in the observed topological metrics. Each point corresponds to the average of 500 random realisations; transparent bands to the 25 – 75 percentile band.

IV. APPLICATION: PROPAGATION OF DELAYS IN AIR TRANSPORT NETWORKS

As an example of the impact that missing and extreme values can have in the calculation of causality relationships, I here revisit the analysis on the propagation of delays in air transport proposed in [32]. The characterisation of delay propagation is one of the most important research topics in air transport management, due to delays' negative implications in the cost-efficiency [43], safety [44], and environment footprint [45] of air transportation. To illustrate, the Federal Aviation Administration estimates that US flight delays cost airlines $22bn yearly. Additionally, 1 minute of ground delay implies between 1 to 4 kg of fuel consumption and between 3 to 12 kg of CO_2 emissions, one order of magnitude higher in the case of airborne delay [45].

While delays have mostly been studied through large-scale simulations and models [46]–[51], a new approach has recently been proposed, based on the reconstruction of delay functional networks [32], [52]–[54]. Inspired on the way information transmission is represented in neuroscience [55], [56], airports are mapped into nodes of a network, with pair of them connected whenever a (statistically significant) causality relationship between their delay evolution is detected. In other words, a propagation process is assumed to be taking place between airports A and B whenever an increase in the average delays observed in A is usually followed by an increase in B. The main advantage of this approach is that no a priori assumptions are needed, e.g. on aircraft or passengers connectivity; the analysis instead only relies on observable time series. On the other hand, it has to be noted that, if the Granger causality test is used, what detected is not causality in general but rather a *predictive causality* [3].

These initial studies on delay functional networks [32], [52]–[54] were based on time series describing, for each airport, the hourly average delay of arriving (or departing) flights, without taking into account their incompleteness. This may be due to two main causes. On one hand, many airports do not operate around the clock, such that some hours can have no operations associated to them, and the corresponding average delay would be zero. Yet, this zero is not equivalent to having no delays, but instead represents a missing value. In other words, we cannot know what would be the expected delay at the airport, would a flight have landed at that time. Similarly, an aircraft arriving late in a period with few or no other operations can substantially change the average delay for that period; once again, this does not correspond to a true status in which all flights are delayed, but just represents a spurious value.

A. DATA AND THEIR PREPROCESSING

The analysis here proposed is based on the same data set described in [32], which includes time series of average delays at the 50 largest European airports. These time series have been obtained by analysing aircraft trajectories included in the Flight Trajectory (ALL-FT+) data set provided by the EUROCONTROL's PRISME group. All flights crossing the European airspace are described through their planned and executed trajectories, with positions reported on average every 2 minutes. The data set covers the period from 1^{st} March to the 31^{st} December 2011, including a total of $10.3 \cdot 10^6$ flights. Only flights landing at the 50 busiest European airports (in terms of number of operations) have further been processed.

A time series has been extracted for each airport, representing the average hourly delay of arriving flights. Delays are here calculated as the difference between actual and planned landing time, which correspond to the delays experienced by passengers. Negative delays, i.e. when an aircraft arrived before its scheduled time, have not been deleted. Missing and extreme data have been identified as respectively those time windows in which no aircraft has landed (12.85% of the data), and in which the absolute value of the average delay was larger than one hour (2.68%). The choice of this last threshold is a subjective one, as no rule exists to define what an abnormal delay is. Still, an average delay larger than one hour implies a major disruption of the system, and is a quite unusual event. The frequency of missing and extreme values in each airport was weakly correlated, albeit in a non statistically significant way, to the ranking of the airport, with a Spearman's rank correlation of respectively 0.32 (p-value of .025) and -0.15 (p-value of 0.299); smaller airports thus have more missing values but fewer extreme values.

As a final step, it is worth noting that the resulting time series are not stationary, as delays are usually correlated to traffic volumes: they are higher during peak hours, week days and the summer. These peaks occurring at the same time at different airports may introduce spurious correlations, and hence spurious causalities. In order to make the time series stationary, a detrend process has been performed, based on subtracting the average delay observed in the same day of the week in the two previous and following weeks, at the same hour:

$$\bar{d}(t) = d(t) - \frac{1}{4} \sum_{i \in \{-2,-1,1,2\}} d(t + 168i), \qquad (12)$$

$d(t)$ being the original time series at time t, and $\bar{d}(t)$ the final time series. Note that the factor of 168 corresponds to the number of hours composing one week, and i spans between two weeks in the past to two weeks in the future. According to this definition, $\bar{d}(t)$ represents the difference (or the *surprise*) between the observed and the expected (historical) delay. Information about the future dynamics of the system should not generally be used to detrend the time series, as this could bias the Granger test - which is actually based on comparing past and future dynamics. This is nevertheless not a problem in the present analysis, as the maximum delay considered (i.e. the maximum lag in the ARMA model) is of eight hours, and the detrend process uses information of one week in the future.

Functional networks are finally reconstructed by evaluating the Granger causality test between all pairs of airports, and by creating a link between the corresponding nodes whenever a significant p-value is obtained ($\alpha = 0.01$ with a Šidák correction for multiple comparisons).

B. SCENARIOS DEFINITION

Following the results obtained in Sec. III, three different scenarios are here considered.

The first one, labelled *raw data* in the following figures and tables, involves no missing and extreme values handling. This is expected to yield a low number of significant causality relations and, most importantly, a biased network topology - for instance a biased assessment of which airports are most important from the delay propagation view-point. Note that this is the customary approach, as e.g. in [32], [52]–[54].

The second and third scenarios respectively involve substituting wrong values for the median of the corresponding distribution (labelled *median*), and weighting the linear regression model (labelled *weighting*). Both methods are expected to recover a larger number of relationships, with the latter potentially yielding a better view of the real topology.

C. RESULTS: DELAY PROPAGATION NETWORKS

The results of applying this functional network reconstruction process are shown in Fig. 10 and Tab. 1, for the three considered scenarios. The network reconstructed by weighting the regression model is the densest of the three, as should be expected given the efficiency of this strategy in recovering lost causality connections. While the efficiency (a metric describing how efficiently the network exchanges information [37]) is almost constant in the three cases, this is not true for the transitivity (density of triangles [36]) and assortativity (correlation between the degrees of connected nodes [38]). Specifically, in the latter case the network changes from assortative (positive metric value, nodes with large number of links tend to connect between themselves) to disassortative (negative metric value, nodes with high number of links tend to avoid themselves). In order words, airports mostly responsible for the propagation of the delays move from being connected to each other, to becoming the center of separate communities. Note that this is a major structural change and not just the result of the higher link density; specifically, if links are deleted from the weighted model network at random, until reaching the link density corresponding to the raw data case, the resulting assortativity is still negative and equal to -0.179 ± 0.062 (average and standard deviation over 100 random realisations). The differences in the obtained connectivity is also graphically represented in Fig. 10, both for the full network (top panels) and for the sub-network composed by the ten most connected airports (central panels). As previously seen, there is a change in the structure created by these highly-connected (in terms of delay propagation) airports, which also reflects in the highly skewed degree distribution of the *median* and *weighted* cases (bottom panels).

As a final analysis, Tab. 2 reports, for the three cases, a list of the five airports most central in the propagation of delays. The importance of each airport has been estimated through the well-known *eigenvector centrality*, a metric which assigns

TABLE 1. Values of four topological metrics (see main text for definitions) describing the structure of the delay propagation networks, as calculated through raw data, median and weighted model pre-processing. Numbers in parenthesis correspond to the Z-Score of the metric, i.e. the distance between the observed metric and what expected in an ensemble of random equivalent (same number of nodes and links) graphs [39].

Topological metric	Raw data	Median	Weighted
Link density	0.094	0.158	0.173
Transitivity	0.117 (0.969)	0.364 (12.7)	0.261 (5.00)
Efficiency	0.419 (−0.454)	0.496 (−2.92)	0.561 (0.540)
Assortativity	0.188	0.0253	−0.208

TABLE 2. Ranking of the five most important nodes, as measured through the corresponding eigenvector centrality [57], [58], calculated through raw data, median and weighted model pre-processing. Numbers in parenthesis correspond to the value of the centrality, normalised such that the most influential node has a centrality of one. Airport are identified by the corresponding 4-letters ICAO codes.

Ranking	Raw data	Median	Weighted
1st	EDDL (1.0)	ENGM (1.0)	EDDM (1.0)
2nd	LEPA (0.624)	LFPO (0.945)	LEPA (0.886)
3rd	EGKK (0.617)	LEBL (0.864)	EHAM (0.669)
4th	EKCH (0.563)	LGAV (0.729)	EDDL (0.511)
5th	EHAM (0.247)	EHAM (0.715)	EGLL (0.490)

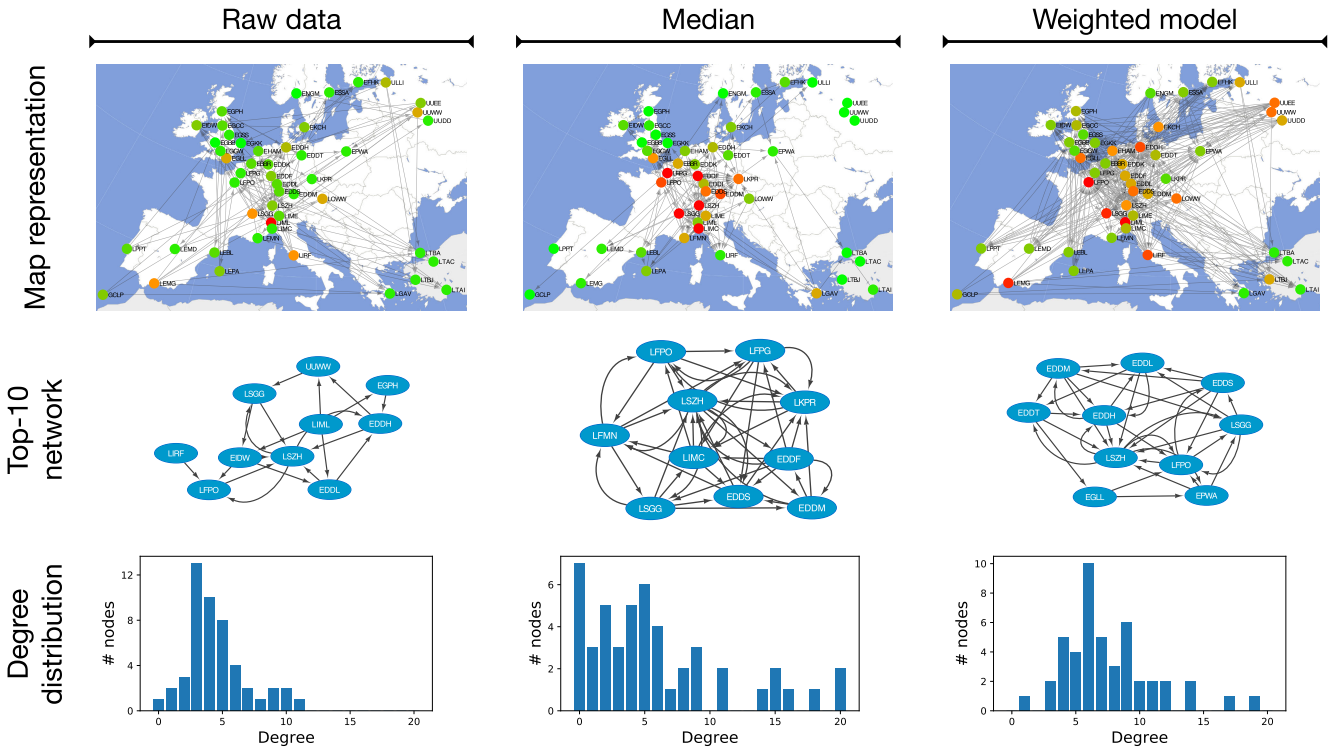

FIGURE 10. Delay propagation networks, as calculated using raw data (left panels), median (central panels) and weighted model (right panels) pre-processing. Top panels report the graphical representation of the full networks; node colours indicate their number of outbound connections, from green (loosely connected airports) to red (densely connected airports). Note that, for the sake of clarity, airport positions are not exact - see, for instance, the case of GCLP (Gran Canaria Airport). Central panels depict the sub-network created by the top-10 most connected airports. Finally, bottom panels report the degree distribution of nodes, i.e. the histogram of nodes' number of links.

to each node an importance proportional to the sum of the importance of neighbouring nodes [57], [58]. Mathematically, suppose the centrality of node i is represented by c_i; and its connectivity by the elements of the adjacency matrix $\mathcal{A}$, as introduced in Eq. 3, such that $a_{i,j} = 1$ if a link exists connecting node i with node j. The centrality of node i is then defined as:

$$c_i = \frac{1}{\lambda} \sum_j a_{i,j} c_j, \tag{13}$$

with λ being a constant. Eq. 13 can be rewritten in vector notation as the eigenvector equation $\mathcal{A}c = \lambda c$, hence the name of eigenvector centrality. In the context of air transport, this metric is measuring how instrumental is each airport in the propagation of delays. In other words, an airport in a central position is more likely to spread the delay it generates to other airports; but also to receive external delays, thus effectively acting as a propagation intermediary. This centrality thus also indirectly measures what would be the benefit if resources aimed at stopping the propagation were assigned to each airport. The centrality measure included in Tab. 2 has been normalised, such that the most influential node has a centrality of 1. It can be appreciated that different data pre-processing strategies correspond to different airport rankings. Specifically, the most central airport when using the raw data (EDDL, Düsseldorf Airport) loses half of its importance in a weighted model; and airports like EDDM (Munich Airport) and EHAM (Amsterdam Airport Schiphol) raise in importance.

This example illustrates the importance of correctly handling missing and extreme values when analysing real data, and also the risks one may otherwise incur. A natural way of exploiting the functional delay networks here reconstructed is to identify the most central airports, for then increase the resources there available. These resources could be physical, as e.g. new runways or more air traffic controllers; but also virtual, like granting priority to flights at them landing [59], [60]. In both cases, delays at these airports will be reduced, causing a disruption in the propagation process and an improvement of the overall dynamics.

The key point is the correct identification of the most central airports, especially considering that it would be difficult at best to validate the results of this functional analysis without modifying the real system. When missing and extreme values are not accounted for, the resulting centralities could be biased. The analysis may then, for instance, conclude that small airports are more central: this is the case in Tab. 2 of Düsseldorf Airport (EDDL) and Copenhagen Airport (EKCH), or of the main conclusions drawn in [53]. The reality may nevertheless be different, with a correction of missing values leading to an increase centrality of Munich Airport (EDDM), Amsterdam Airport Schiphol (EHAM) and

Heathrow Airport (EGLL), that is, of the largest airports in Europe. In short, the practitioner should be aware that results obtained without a proper handling of missing and extreme values may be highly unreliable.

V. DISCUSSION AND CONCLUSION

Granger causality is one of the most famous, and possibly the most used causality test in science and engineering; still, the study of its behaviour in the presence of missing and extreme values has not received a significant attention, especially the case in which these values are distributed randomly throughout the time series. This is an important issue as, on one hand, erroneous values are a common denominator in many real-world applications [24]; and, on the other hand, the Granger test can yield wrong results when even a small percentage of the data is modified - see Fig. 1. Missing values have an even more profound effect when the Granger test is used to reconstruct functional network representations, as the resulting bias is a function of how missing values are distributed and what topological property is analysed - see Fig. 2.

This contribution addresses this problem by comparing three ways for data pre-processing, two based on standard value substitution, and a novel one based on redefining the underlying linear model to disregard wrong values. Additionally, both dyadic and network-based causality relations have been considered. While a full discussion of the results is included in Sec. III-D, the synthesis is that the weighted model is the best option, followed by substituting wrong values by the median of the distribution.

In order to test these three strategies in a real-world situation, Sec. IV revisits the problem of detecting delay propagation in an air transport system by means of functional networks, reconstructed by detecting causality relationships between observed delay time series. While such analysis is not new, previous studies [32], [52]–[54] have neglected the presence of false values, mostly generated by the absence of landing aircraft in some time windows. As shown in Tabs. 1 and 2, compensating for those wrong values substantially change the observed propagation structure. Specifically, hub airports (here understood as those airports most responsible for the propagation) change from being connected between them, to form independent communities; and the identity of those airports also substantially change. This example illustrates the importance of a correct handling of wrong values, as even a small fraction of them can radically change the observed results.

It is important to highlight that the guidelines presented in Sec. III-D should not substitute the complete study, as the researcher has to take in consideration the idiosyncrasies of the problem at hand - and of its associated data. For instance, the strength of the causality relation (represented by γ in the synthetic models) is usually not known in a real-world problem, and has to be estimated. Also, the way missing and wrong values are encoded may affect the results of data pre-processing strategies, as shown in Fig. 9. Finally, the same definition of what a missing or an extreme value is depends on the problem at hand - for instance, missing values can be clearly labelled, or may appear as normal values as in the case of zero delays. In synthesis, no data pre-processing strategy can substitute the judgement of the researcher and his/her knowledge of the problem at hand.

REFERENCES

[1] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica, J. Econ. Soc.*, vol. 37, no. 3, pp. 424–438, Aug. 1969.
[2] N. Wiener, "The theory of prediction," in *Modern Mathematics for Engineers*. New York, NY, USA: McGraw-Hill, 1956.
[3] F. X. Diebold, *Elements of Forecasting*. Tokha Saraswati, Nepal: South-Western College, 1998.
[4] W. Joerding, "Economic growth and defense spending: Granger causality," *J. Develop. Econ.*, vol. 21, no. 1, pp. 35–40, 1986.
[5] L. Kónya, "Exports and growth: Granger causality analysis on OECD countries with a panel data approach," *Econ. Model.*, vol. 23, no. 6, pp. 978–992, Dec. 2006.
[6] C. Diks and V. Panchenko, "A new statistic and practical guidelines for nonparametric granger causality testing," *J. Econ. Dyn. Control*, vol. 30, nos. 9–10, pp. 1647–1669, Sep. 2006.
[7] P. K. Narayan and R. Smyth, "Multivariate granger causality between electricity consumption, exports and GDP: Evidence from a panel of middle eastern countries," *Energy Policy*, vol. 37, no. 1, pp. 229–236, Jan. 2009.
[8] T. Yuan and S. J. Qin, "Root cause diagnosis of plant-wide oscillations using granger causality," *J. Process Control*, vol. 24, no. 2, pp. 450–459, Feb. 2014.
[9] J. R. Freeman, "Granger causality and the times series analysis of political relationships," *Amer. J. Political Sci.*, vol. 27, no. 2, pp. 327–358, May 1983.
[10] S. Kleinberg and G. Hripcsak, "A review of causal inference for biomedical informatics," *J. Biomed. Informat.*, vol. 44, no. 6, pp. 1102–1112, Dec. 2011.
[11] S. L. Bressler and A. K. Seth, "Wiener–Granger causality: A well established methodology," *NeuroImage*, vol. 58, pp. 323–329, Sep. 2010.
[12] A. K. Seth, A. B. Barrett, and L. Barnett, "Granger causality analysis in neuroscience and neuroimaging," *J. Neurosci.*, vol. 35, no. 8, pp. 3293–3297, Feb. 2015.
[13] P. A. Stokes and P. L. Purdon, "A study of problems encountered in granger causality analysis from a neuroscience perspective," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 34, pp. E7063–E7072, Aug. 2017.
[14] L. Barnett and A. K. Seth, "Granger causality for state-space models," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 91, no. 4, Apr. 2015, Art. no. 040101.
[15] L. Barnett and A. K. Seth, "Detectability of granger causality for subsampled continuous-time neurophysiological processes," *J. Neurosci. Methods*, vol. 275, pp. 93–121, Jan. 2017.
[16] M. Zanin and D. Papo, "Detecting switching and intermittent causalities in time series," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 27, no. 4, Apr. 2017, Art. no. 047403.
[17] S. Stramaglia, T. Scagliarini, Y. Antonacci, and L. Faes, "Local granger causality," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 103, no. 2, Feb. 2021, Art. no. L020102.
[18] H. Nalatore, M. Ding, and G. Rangarajan, "Mitigating the effects of measurement noise on Granger causality," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 75, no. 3, Mar. 2007, Art. no. 031123.
[19] H. Nalatore, N. Sasikumar, and G. Rangarajan, "Effect of measurement noise on Granger causality," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 90, no. 6, Dec. 2014, Art. no. 062127.
[20] M. Vinck, L. Huurdeman, C. A. Bosman, P. Fries, F. P. Battaglia, C. M. A. Pennartz, and P. H. Tiesinga, "How to detect the Granger-causal flow direction in the presence of additive noise?" *NeuroImage*, vol. 108, pp. 301–318, Mar. 2015.
[21] B. D. O. Anderson, M. Deistler, and J. Dufour, "On the sensitivity of Granger causality to errors-in-variables, linear transformations and subsampling," *J. Time Ser. Anal.*, vol. 40, no. 1, pp. 102–123, Jan. 2019.
[22] M. T. Bahadori and Y. Liu, "Granger causality analysis in irregular time series," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, Apr. 2012, pp. 660–671.

[23] H. Elsegai, "Granger-causality inference in the presence of gaps: An equidistant missing-data problem for non-synchronous recorded time series data," *Phys. A, Stat. Mech. Appl.*, vol. 523, pp. 839–851, Jun. 2019.
[24] T. D. Pigott, "A review of methods for missing data," *Educ. Res. Eval.*, vol. 7, no. 4, pp. 353–383, 2001.
[25] M. N. Norazian, Y. A. Shukri, and R. N. Azam, "Estimation of missing values in air pollution data using single imputation techniques," *Science Asia*, vol. 34, no. 2, pp. 341–345, 2008.
[26] S. Arndt, T. Cizadlo, N. C. Andreasen, D. Heckel, S. Gold, and D. S. O'Leary, "Tests for comparing images based on randomization and permutation methods," *J. Cerebral Blood Flow Metabolism*, vol. 16, no. 6, pp. 1271–1279, Nov. 1996.
[27] P. M. T. Broersen, S. de Waele, and R. Bos, "Application of autoregressive spectral analysis to missing data problems," *IEEE Trans. Instrum. Meas.*, vol. 53, no. 4, pp. 981–986, Aug. 2004.
[28] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 28, no. 7, Jul. 2018, Art. no. 075310.
[29] M. Zanin, D. Papo, P. A. Sousa, E. Menasalvas, A. Nicchi, E. Kubik, and S. Boccaletti, "Combining complex networks and data mining: Why and how," *Phys. Rep.*, vol. 635, pp. 1–44, May 2016.
[30] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, Mar. 2001.
[31] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.
[32] M. Zanin, "Can we neglect the multi-layer structure of functional networks?" *Phys. A, Stat. Mech. Appl.*, vol. 430, pp. 184–192, Jul. 2015.
[33] H. Lütkepohl and H.-E. Reimers, "Granger-causality in cointegrated VAR processes the case of the term structure," *Econ. Lett.*, vol. 40, no. 3, pp. 263–268, Nov. 1992.
[34] A. B. Barrett, L. Barnett, and A. K. Seth, "Multivariate Granger causality and generalized variance," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 81, no. 4, Apr. 2010, Art. no. 041907.
[35] L. D. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: A survey of measurements," *Adv. Phys.*, vol. 56, no. 1, pp. 167–242, Jan. 2007.
[36] M. Á. Serrano and M. Boguñá, "Clustering in complex networks. I. General formalism," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 74, no. 5, Nov. 2006, Art. no. 056114.
[37] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Phys. Rev. Lett.*, vol. 87, no. 19, Oct. 2001, Art. no. 198701.
[38] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett.*, vol. 89, no. 20, Oct. 2002, Art. no. 208701.
[39] M. Zanin, X. Sun, and S. Wandelt, "Studying the topology of transportation systems through complex networks: Handle with care," *J. Adv. Transp.*, vol. 2018, pp. 1–17, Aug. 2018.
[40] A. Roebroeck, E. Formisano, and R. Goebel, "Mapping directed influence over the brain using granger causality and fMRI," *NeuroImage*, vol. 25, no. 1, pp. 230–242, Mar. 2005.
[41] J. C. Rajapakse and J. Zhou, "Learning effective brain connectivity with dynamic Bayesian networks," *NeuroImage*, vol. 37, no. 3, pp. 749–760, Sep. 2007.
[42] S. H. Davis and M. N. Roppo, "Coupled lorenz oscillators," *Phys. D, Nonlinear Phenomena*, vol. 24, nos. 1–3, pp. 226–242, Jan. 1987.
[43] A. J. Cook and G. Tanner, "European airline delay cost reference values," EUROCONTROL Perform. Rev. Unit, Brussels, Belgium, Tech. Rep., 2011. [Online]. Available: https://westminsterresearch.westminster.ac.uk/item/8zxq0/european-airline-delay-cost-reference-values
[44] D. Duytschaever, "The development and implementation of the EUROCONTROL central air traffic flow management unit (CFMU)," *J. Navigat.*, vol. 46, no. 3, pp. 343–352, Sep. 1993.
[45] S. Carlier, I. De Lépinay, J.-C. Hustache, and F. Jelinek, "Environmental impact of air traffic flow management delays," in *Proc. 7th USA/Eur. Air Traffic Manage. Res. Develop. Seminar (ATM)*, vol. 2, 2007, p. 16.
[46] K. F. Abdelghany, S. S. Shah, S. Raina, and A. F. Abdelghany, "A model for projecting flight delays during irregular operation conditions," *J. Air Transp. Manage.*, vol. 10, no. 6, pp. 385–394, Nov. 2004.
[47] M. Janić, "Modeling the large scale disruptions of an airline network," *J. Transp. Eng.*, vol. 131, no. 4, pp. 249–260, Apr. 2005.
[48] M. Jetzki, "The propagation of air transport delays in Europe," Ph.D. dissertation, Dept. Airport Air Transp. Res. RWTH, Aachen Univ., Aachen, Germany, 2009.
[49] P. Fleurquin, J. J. Ramasco, and V. M. Eguiluz, "Systemic delay propagation in the US airport network," *Sci. Rep.*, vol. 3, no. 1, pp. 1–6, Dec. 2013.
[50] N. Kafle and B. Zou, "Modeling flight delay propagation: A new analytical-econometric approach," *Transp. Res. B, Methodol.*, vol. 93, pp. 520–542, Nov. 2016.
[51] H. Zhang, W. Wu, S. Zhang, and F. Witlox, "Simulation analysis on flight delay propagation under different network configurations," *IEEE Access*, vol. 8, pp. 103236–103244, 2020.
[52] M. Zanin, S. Belkoura, and Y. Zhu, "Network analysis of chinese air transport delay propagation," *Chin. J. Aeronaut.*, vol. 30, no. 2, pp. 491–499, Apr. 2017.
[53] W.-B. Du, M.-Y. Zhang, Y. Zhang, X.-B. Cao, and J. Zhang, "Delay causality network in air transport systems," *Transp. Res. E, Logistics Transp. Rev.*, vol. 118, pp. 466–476, Oct. 2018.
[54] P. Mazzarisi, S. Zaoli, F. Lillo, L. Delgado, and G. Gurtner, "New centrality and causality metrics assessing air traffic network interactions," *J. Air Transp. Manage.*, vol. 85, Jun. 2020, Art. no. 101801.
[55] E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems," *Nature Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, Mar. 2009.
[56] D. S. Bassett and E. T. Bullmore, "Human brain networks in health and disease," *Current Opinion Neurol.*, vol. 22, no. 4, p. 340, 2009.
[57] P. Bonacich, "Power and centrality: A family of measures," *Amer. J. Sociol.*, vol. 92, no. 5, pp. 1170–1182, Mar. 1987.
[58] P. Bonacich, "Some unique properties of eigenvector centrality," *Social Netw.*, vol. 29, no. 4, pp. 555–564, Oct. 2007.
[59] M. Sama, A. D'Ariano, P. D'Ariano, and D. Pacciarelli, "Scheduling models for optimal aircraft traffic control at busy airports: Tardiness, priorities, equity and violations considerations," *Omega*, vol. 67, pp. 81–98, Mar. 2017.
[60] E. Grunewald, F. Knabe, F. Rudolph, and M. Schultz, "Priority rules as a concept for the usage of scarce airport capacity," *Transp. Res. Procedia*, vol. 27, pp. 1146–1153, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146517309341

MASSIMILIANO ZANIN was born in Verona, Italy, in 1982. He received the Ph.D. degree in computer engineering from the Universidade Nova de Lisboa, Portugal, in 2014.

He is currently a Researcher with the Institute for Cross-Disciplinary Physics and Complex Systems, Palma de Mallorca, Spain. He is also the PI of the ERC StG ARCTIC, on the analysis and modeling of delay propagation in air transport. He has participated in several European competitive research projects. Throughout his career he has published more than 80 articles in international journals, and more than 50 contributions in conference papers, reaching an H-index of 25. His research interests include complex networks and data science, both from a theoretical perspective and through their application to several real-world problems.