

POLITECNICO DI TORINO

Corso di Laurea
in Matematica per l'Ingegneria

Tesi di Laurea magistrale

Complex Network-based Analysis of
Climate Series



Relatori

Luca Ridolfi
Stefania Scarsoglio
Pietro S. Salizzoni

(École Centrale de Lyon)

Candidata

Camilla Viazzo

Marzo 2022

Acknowledgements

First and foremost, I would like to thank my tutors for the opportunity to work on this thesis, as well as Luca Mercalli for providing us with all the data for the series of Turin. A special thanks goes to prof. Ridolfi for his guidance and support.

I am also extremely grateful to all the people that accompanied me in these months regardless of the distance, in particular Erwin.

Last but not least, I would like to thank Yun for everything.

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 4 |
| 2 | Climatic series | 5 |
| 2.1 | Preprocessing | 6 |
| 2.1.1 | Homogenization | 6 |
| 2.1.2 | Break detection methods | 7 |
| 2.2 | Meteclimatic indicators | 12 |
| 2.3 | Time series reversibility | 13 |
| 3 | Complex networks | 15 |
| 3.1 | Elements of graph theory | 15 |
| 3.1.1 | Measures and indices | 16 |
| 3.2 | Visibility Graphs | 16 |
| 3.2.1 | Irreversibility and Kullback-Leibler divergence | 20 |
| 3.2.2 | Motifs and system dynamics | 22 |
| 4 | Case study 1: Turin 1753-2020 | 26 |
| 4.1 | Series description | 26 |
| 4.1.1 | Homogenization | 28 |
| 4.2 | Preliminary analysis | 32 |
| 4.3 | Visibility graphs | 38 |
| 4.3.1 | Degree metrics | 38 |
| 4.3.2 | Motif detection | 46 |
| 5 | Case study 2: Prague | 48 |
| 5.1 | Series description | 48 |
| 5.1.1 | Homogenization | 48 |
| 5.1.2 | Missing values | 50 |
| 5.1.3 | Preliminary analysis | 50 |
| 5.2 | Visibility graphs | 54 |

| | | |
|----------|-----------------------------------|-----------|
| 5.2.1 | Degree metrics | 54 |
| 5.2.2 | Assortativity | 55 |
| 5.2.3 | Time reversibility | 59 |
| 5.2.4 | Motif detection | 60 |
| 6 | Case study 3: Bologna | 62 |
| 6.1 | Series description | 62 |
| 6.1.1 | Preliminary analysis | 62 |
| 6.2 | Visibility graphs | 67 |
| 6.2.1 | Degree metrics | 67 |
| 6.2.2 | Assortativity | 69 |
| 6.2.3 | Time reversibility | 69 |
| 6.2.4 | Motif detection | 73 |
| 7 | Discussion and Conclusions | 75 |
| 7.1 | Case study comparisons | 75 |
| 7.2 | Final remarks | 79 |

Chapter 1

Introduction

Long-term instrumental daily air temperature series are fundamental for climate monitoring, climate change detection and attribution, climate modelling and to assist climate action and adaption policies [16]. During the reconstruction of these series, it is crucial to adjust for missing values and inhomogeneities caused by nonclimatic factors and accumulated over time; this is generally achieved by combining statistical analysis with the available metadata and historical records available.

In recent years a bridge between time series analysis and complex networks has been proposed to characterize and model the macroscopic and microscopic structure of complex systems in nature, technology, and society [30], bypassing some of the challenges that characterize traditional approaches. A rather intuitive method to handle scalar time series is the visibility graph, introduced by Lacasa et al. [12]; it can be used to evaluate time irreversibility [14] and perform robust discrimination between different types of complex dynamics through motif analysis [10],[28].

The thesis can be divided in two sections: the first one (corresponding to chapters 2 and 3) gives a brief introduction to climate analysis and lays the theoretical foundation for the analysis and the results discusses in the second one. In particular section 2.1.1 introduces the homogenization techniques that will be further discussed in the presentation of the case studies, and chapter 3 introduces the visibility graphs. The main focus is the application of the visibility graphs to three long-term climatic series, presented in chapters 4-6.

Chapter 2

Climatic series

Climate change refers to long-term shifts in temperatures and weather patterns. These shifts may be attributed to natural phenomena, such as variations in the solar cycle, but since the 1800s human activities have been the main driver of climate change, primarily due to burning fossil fuels like coal, oil and gas. The prevalent anthropogenic effect is the warming caused by the increase in greenhouse gases [IPCC 2007].

Volcanic eruptions

Volcanic activity is considered to be the primary cause of interdecadal variability of the climate [17]. The impact of large-scale volcanic eruptions on the climate is caused by the emission in the stratosphere of huge amounts of dust and sulfur dioxide, that are transported by the wind on the entire hemisphere - or the planet, if the volcano is located in the tropical region. The aerosol changes the energy flux in the atmosphere by diffusing part of the incident solar radiation, without interfering with the radiation emitted from the Earth: this phenomenon, called *radiative forcing*, results in an increase in temperature in the stratosphere and a cooling of the surface of the Earth and the low troposphere. The decrease in temperature is maximum in the first year after the eruption, and it lasts generally between one and three years, after which the aerosol falls again in the troposphere and on the ground through precipitations.

An overview of the major volcanic eruptions may help to interpret anomalies in the average yearly temperatures of the series considered. The eruption of the Tambora in Indonesia in 1815, for example, was the greatest eruption in the modern age, and caused 1816 to be the “Year Without a Summer” in most of the North hemisphere; the combination of low temperatures, huge

storms and abnormal rainfall resulted in an agricultural disaster. To give a perspective on the intensity of the phenomenon, its radiative forcing is estimated to be $-4W/m^2$ [29], whereas the radiative forcing associated with the increase in greenhouse gases since the Pre-Industrial Era - net of aerosol and clouds - is estimated to be between $+1.5W/m^2$ and $+2.8W/m^2$.

2.1 Preprocessing

The first consistency checks that have to be run to detect potential errors in the data are internal, temporal, spatial and summarization [27]. The first one checks for consistency with definitions - e.g. the maximum value has to be always greater or equal than the corresponding minimum value -, physical bounds - e.g. precipitations cannot be negative - and on a deeper level relies on the physical relationships among climatological elements. The temporal consistency check evaluates the relationship of a data point with the preceding and successive one, flagging changes suspiciously far from the expected amount. Similarly, the spatial consistency test compares neighbouring observations within a climatologically similar area recorded at the same time and flags the outliers. Summarization tests can be used for example to cross-check data summaries with different time aggregation. The entries flagged as errors lead to gaps in the data, and their value is estimated during the process of homogenization of the series.

2.1.1 Homogenization

An issue particularly prevalent in long climatic series is the presence of inhomogeneities, caused mainly by the relocation of meteorological stations, replacement or recalibration of instrumentation, and a change in the surrounding environment, due to urban expansion for example. The induced shifts (or breaks) often have the same magnitude of the climate signal, such as long-term variations, trends or cycles, and might lead to wrong conclusions about the evolution of the climate[3]. In order to obtain a *homogeneous* climate series, i.e. that is only influenced by the variations in climate, it is necessary to perform gap completion, break detection and correction.

The availability of metadata can be crucial to correctly identify breaks, estimate corrections and validate results, but especially for older records it may be incomplete. In section 4.1 and 5.1 two possible homogenization approaches are presented: in the first one the metadata available is taken into account, whereas the second application is blind to metadata and operates

automatically. This may be preferable when the dataset considered is particularly big (e.g. all the major European meteorological stations) and the metadata is insufficient. Here is a brief introduction to the methods that will be used for the homogenization of the series in this case study.

EM algorithm

The Expectation Maximization algorithm, proposed by Dempster et al. (1977), is a two-step iterative method that can be used to estimate missing data in time series[7].

Let x and z the observed and missing data respectively, θ be the unknown parameter vector and θ_n its estimate at iteration n . The E-step calculates the conditional expectations of missing data given observed data and estimates of model parameters as:

$$Q(\theta|\theta_n) = \mathbb{E}_{Z|x,\theta_n}[\log L(\theta; x, z)], \quad (2.1)$$

where $L(\theta; x, z)$ is the likelihood function. The M-step then finds the estimates of the model parameter that maximize the complete-data log likelihood function from the E-step:

$$\theta^* = \arg_{\theta} \max Q(\theta|\theta_n), \quad (2.2)$$

and the process is iterated until convergence.

2.1.2 Break detection methods

The break detection methods can be classified in three categories: likelihood-based, linear regression-based, and nonparametric. They generally rely on the assumption that the difference between the series under study and a reference series is fairly constant in time [3], and most of the breaks (also referred to as shifts) are step-like changes which typically alter only the average. Given $K - 1$ changepoints occurring at times $\{\tau_1, \dots, \tau_{K-1}\}$, with K unknown, the time series can be divided into K homogeneous segments in terms of statistical features, such as the mean; in this sense changepoint detection and time series segmentation are equivalent problems.

A brief overview of some of the most common changepoint detection methods is presented.

Prodige: Caussinus and Mestre

The Caussinus and Lyazrhi's[15] procedure for break detection is based on the pairwise comparison of the test series with a set of reference series from the same climatic area. If a changepoint remains constant through the comparisons, it can be attributed to the test series; in this way it is possible to distinguish the breaks detected due to an unreliable reference series from those of the test series.

Let X be a matrix with X_{ij} the observation at time $i \in \{1, \dots, n\}$ of the series at station $j \in \{1, \dots, p\}$; given the series j , let k_j and l_j the number of changepoints and outliers, and $K_j = (\{0, \tau_{1,j}, \dots, \tau_{k_j,j}, n\}, \{\delta_{1,j}, \dots, \delta_{l_j,j}\})$ their respective positions¹. Let $L_{jh} = [\tau_{h-1,j} + 1, \tau_{h,j}]$ the *level* h : by definition, each level is an homogeneous subperiod of the series. The observations are assumed to be the sum of a climate effect μ_i at time i , a station effect ν_{jh} of station j for the level L_{jh} and random white noise; the station effect is piecewise constant between two shifts, and - conditionally to the climate signal - the disturbances can be considered independent [3] .

The data is described by the linear model

$$\begin{aligned} \mathbb{E}(X_{ij}) &= \mu_i + \nu_{jh(i,j)} \\ \text{Var}(X) &= \sigma^2 I_{np}, \end{aligned} \tag{2.3}$$

where the notation $h(i, j)$ remarks that level h for observation X_{ij} depends both on time i and station j . An additional parameter is added to the mean if the data indexed by (i, j) are outliers. The parameters are identified by $\sum_{i=1}^n \mu_i = 0$.

Let K be the union of the positions of changepoints and outliers of all series, and define k and l as the total number of changepoints and outliers respectively; assuming normality of the observations, the penalized log-likelihood procedure is

$$\text{select } H_{K^*} \quad \text{s.t. } K^* = \arg \min_K C_K(X), \tag{2.4}$$

¹In this simplified notation, $\tau_{0,j} = 0$ and $\tau_{k_j+1,j} = n$

where $C_\emptyset(X) = 0$ and

$$C_K(X) = \ln \left[1 - \frac{\sum_{j=1}^p \sum_{i=1}^n [(\hat{\mu}_i^K + \hat{\nu}_{jh(i,j)}^K)^2 - (\hat{\mu}_i^\emptyset + \hat{\nu}_j^\emptyset)^2]}{\sum_{j=1}^p \sum_{i=1}^n [X_{ij} - (\hat{\mu}_i^\emptyset + \hat{\nu}_j^\emptyset)]^2} \right] \\ + \frac{2(k+l)}{np - m - p - n + 1} \ln(np - m),$$

where m is the number of missing values, $\hat{\mu}_i^\emptyset$ and $\hat{\nu}_j^\emptyset$ are the least square estimates under the null hypothesis $K = \emptyset$ and $\hat{\mu}_i^K$ and $\hat{\nu}_{jh(i,j)}^K$ are the estimates under any alternative hypothesis H_K .

The imputation of missing data is given by $\hat{X}_{ij} = \hat{\mu}_i + \hat{\nu}_{jh(i,j)}$, and each observation $X_{ij} \in L_{jh}$ ($1 \leq h \leq k_j + 1$) is corrected by

$$X_{ij}^* = X_{ij} - \hat{\nu}_{jh(i,j)}^{K*} + \hat{\nu}_{j,k_j+1}^{K*}. \quad (2.5)$$

It is important to note that the number of hypothesis rises very fast with the length of the series n and the number of accidents $k+l$, rendering a naïve implementation of the model unfeasible. Assuming the normality hypothesis, Mestre [18] proposed a stepwise algorithm that limits the detection of new accidents at every step to one outlier or one or two breaks; in this way the computation time scales quadratically with n . A more general approach to reduce the number of hypothesis to test in (2.4) consists in the preselection of changepoints and outliers: this is done by comparing the series pairwise to identify the most probable breaks, correcting the series and iterating the process two or three times to ensure good results.

RHtest: Wang et al.

The method proposed by Wang et al.[25] focuses on the detection of undocumented shifts in the mean. Differently from the approach discussed in the previous section, the series is assumed to have at most one changepoint (AMOC), and multiple breaks can be detected through a recursive testing algorithm. The changepoint in the series $\{X_t\}_{t=1}^n$ is detected by testing the null hypothesis $H_0 : \{X_t\} \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ against the alternative

$$H_a : \begin{cases} \{X_t\} \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma^2), & t = 1, \dots, k \\ \{X_t\} \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma^2), & t = k+1, \dots, n \end{cases}$$

with $\mu_1 \neq \mu_2$ and $t = k$ the candidate changepoint.

The most probable changepoint is associated with the maximum of a log likelihood ratio, or equivalently[4]

$$T_{max} = \max_{1 \leq k \leq n-1} T(k), \quad (2.6)$$

with

$$\begin{aligned} T(k) &= \frac{1}{\hat{\sigma}_k} \left[\frac{k(n-k)}{n} \right]^{1/2} |\bar{X}_1 - \bar{X}_2|, \\ \bar{X}_1 &= \frac{1}{k} \sum_{t=1}^k X_t, \quad \bar{X}_2 = \frac{1}{n-k} \sum_{t=k+1}^n X_t, \\ \hat{\sigma}_k^2 &= \frac{1}{n-2} \left[\sum_{t=1}^k (X_t - \bar{X}_1)^2 + \sum_{t=k+1}^n (X_t - \bar{X}_2)^2 \right]. \end{aligned}$$

This test is called the maximal (two-sample) t test, and it is equivalent to the standard normal homogeneity test (SNHT) proposed by Alexandersson (1986). Its accuracy decreases significantly for points at the extremities of the time series, as the difference in size of the two samples increases and the false alarm rate (FAR) increases with respect to points in the middle of the series. This phenomenon corresponds to a U-shaped curve for the effective level of significance of the test $\text{FAR}_\alpha(k) \sim k$, with $k = 1, \dots, n-1$ and α the level of significance.

To get a more even FAR, an empirical penalized maximal t test (PMT) is proposed:

$$PT_{max} = \max_{1 \leq k \leq n-1} [P(k)T(k)], \quad (2.7)$$

where $P(k)$ is an empirical penalty function that depends heavily on the series length n (for more details, check Wang et al.[25]). This test tends to overpenalize slightly the test statistic for the end points of very long time series ($n \geq 500$), but in general it evens out considerably the FAR over the series w.r.t. the maximal t test and the SNHT.

GAHMDI: Toreti et al.

The genetic algorithm hidden Markov models for detection of inhomogeneities (GAHMDI) in this framework assumes the discrete time process $\{X_t\}_{t=1}^N$ dependent on the *state process* $\{S_t\}_{t=1}^N$: this is an unobservable process taking values in $\{1, \dots, K\}$ that can be considered as the set of nonclimatic factors

affecting the measured values of the time series. X_t are assumed conditionally independent with a distribution (given S_t) that is Gaussian, with mean μ_{S_t} and variance $\sigma_{S_t}^2$. The state process is a Markov chain with transition matrix \mathbf{P} and initial state distribution $\pi = (\pi_1, \dots, \pi_K)$, where $\pi_i = \mathbb{P}(S_1 = i)$. A left-to-right hidden Markov model (HMM) assumes $\pi = (1, 0, \dots, 0)$ and $p_{i,j} = 0 \quad \forall j < i$, which implies that the process cannot come back to a previous condition²; moreover $p_{i,i} \neq 0$ and $p_{i,i+1} = 1 - p_{i,i} \quad \forall i = 1, \dots, K$. The set of parameters $\lambda = \{K, P, \mu_1, \dots, \mu_K, \sigma_1, \dots, \sigma_K\}$ is estimated from an initial guess by fixing $K \in \{1, \dots, K_{max}\}$ a priori and approximating the model likelihood

$$L(\lambda) = \sum_S \prod_{t=1}^N \frac{\exp \left[-\frac{1}{2} \left(\frac{X_t - \mu_{S_t}}{\sigma_{S_t}} \right)^2 \right]}{\sigma_{S_t} \sqrt{2\pi}} p_{S_{t-1}, S_t}$$

with an expectation-maximization approach, namely the Baum-Welch algorithm [26]. Given the set of estimated parameters $\hat{\lambda}$, the optimal segmentation of $\{X_t\}$ is estimated by the Viterbi algorithm: this is based on the term $\delta_t(i) = \max_{s_1, \dots, s_{t-1}} L_2(s_1, \dots, s_{t-1}, s_t = i, x_1, \dots, x_t | \hat{\lambda})$, that maximizes the conditional likelihood L_2 of the state sequence up to time t and ending in a state equal to i [24].

To get global maxima from the Baum-Welch algorithm, the initial state sequence is estimated by a genetic algorithm (GA) that preserves the HMM structure during the process of crossover and mutation and rejects solutions with any segment less than 4 steps long. If a simplified HMM is used [11], μ_k and σ_k are estimated by the mean and standard deviation, respectively, of all the observations belonging to the state k ; moreover the diagonal elements of \mathbf{P} are equal to the number of time steps without change in the state sequence divided by the total number of time steps, with the exception of $p_{k,k} = 1$. Therefore the evaluation function of the GA is the joint likelihood of the state sequence and the observations:

$$L_1(\lambda) = \prod_{t=1}^N (\sqrt{2\pi} \sigma_{S_t})^{-1} p_{S_{t-1}, S_t} e^{-(x_t - \mu_{S_t})^2 / \sigma_{S_t}^2}, \quad (2.8)$$

²Both this constraint and the Gaussian condition can be relaxed.

with $p_{0,1} = 1$. Imposing $p_{i,i} = p \quad \forall i \leq K - 1$, where p is the first element on the main diagonal of \mathbf{P} , we obtain

$$\begin{aligned} \log L_1(\lambda) = & \sum_{t=1}^N \log \left(\frac{1}{\sqrt{2\pi}\sigma_{s_t}} \right) - \sum_{t=1}^N \frac{(x_t - \mu_{s_t})^2}{2\sigma_{s_t}^2} + (K - 1) \log(1 - p) \\ & + [(N - 1) - (K - 1) - (|C_K| - 1)] \log(p), \end{aligned} \quad (2.9)$$

where $C_K = \{s_t | s_t = K\}$ and $|C_K|$ is the cardinality of the set.

The GAHMDI is applied for $K = 1, \dots, K_{max}$, stopping the procedure when the best state solution has at least one state whose time duration is less than four steps. The optimal number of segments is chosen by minimizing an objective function of the form $-\log(\text{likelihood}) + \text{penalty}$.

2.2 Meteoclimatic indicators

The computation of meteoclimatic indicators provides an intuitive summarization of data and can be helpful to extract and present information about a series. The most common time aggregation is annual, but for certain indicators also monthly or decadal aggregations may be of interest. The quality of the input data can be validated by a weak climatological check and an internal consistency check; the first one verifies if the value of the variable lies between a minimum and maximum acceptance threshold, defined a priori; in Italy, for example, the maximum and minimum temperature values should not be lower than -29°C or higher than 49°C according to the SCIA guidelines[1]. The internal consistency check, instead, takes into consideration multiple variables at the same time point and checks for inconsistencies between their values.

As a result of these checks, the indicator can be paired with a validity flag that is equal to 1 when the indicator is valid, and 0 otherwise. Generally it is sufficient to have at least 75% of valid input data to assign a valid flag, but indicators such as extremes, number of values above/below a threshold, and the cumulative sum of a certain quantity over a period of time are particularly susceptible to gaps in the data, so the threshold of valid data has to be raised to 90%.

Given daily observations of minimum (maximum) temperature, one can compute their mean and standard deviation, find the date of minimum (maximum) and its value, count the number of days above a threshold and the

number of frost days. Moreover these data can be used to calculate the average daily temperature series and the daily thermal excursion.

It may also be of interest to study the *persistence*, i.e. the number of consecutive days with values within a certain range: in the case of temperatures, the intervals (in degree Celsius) that characterize the symbolization of the series are defined by the SCIA [1] as $-10 < T_{max} \leq -5$, $-5 < T_{max} \leq 0$, ..., $35 < T_{max} \leq 40$, $T_{max} > 40$ for maximum temperature and $T_{min} \leq -20$, $-20 < T_{min} \leq -15$, $-15 < T_{min} \leq -10$, ..., $15 < T_{min} \leq 20$ for minimum temperature.

Heat waves

A *heat wave* is a period of prolonged abnormally high surface temperatures relative to those normally expected. According to the World Meteorological Organization (WMO) it is defined as a period of at least 6 consecutive days with maximum temperature above the 90th percentile of that day w.r.t. a 30-year reference period (1981-2010 or 1991-2020). Depending on the application of interest, though, the definition can vary significantly; in many instances it is not even consistent between different countries. In the context of health protection, for example, temperature and humidity data are generally combined to estimate the apparent temperature perceived by the human body and the heat stress [21], in order to devise a heat-health warning system (HHWS) to advise vulnerable sections of the population and prevent hospitalizations.

In the following case studies the major heat waves are detected as the 6 hottest consecutive days of the year, provided that the maximum temperature is always above a threshold of 28°C; each one is also paired with the average maximum temperature over the 6-day period to quantify its intensity. This is a definition used by Mercalli [17] in the analysis of the series of Turin, and it is chosen for its simplicity.

2.3 Time series reversibility

A stationary process is *time reversible* if the joint probability distribution of the forward and backward process are statistically equivalent. More formally, given a time series $\Sigma = \{x_1, x_2, \dots, x_N\}$, and its corresponding backward series $\Sigma^* = \{x_N, x_{N-1}, \dots, x_1\}$, the forward and backward joint distributions

will be $\mathcal{P}_F(N) := P(x_1, x_2, \dots, x_N)$ and $\mathcal{P}_B(N) := P(x_N, x_{N-1}, \dots, x_1)$ respectively; Σ is statistically time reversible if and only if $\mathcal{P}_F(m) \stackrel{d}{=} \mathcal{P}_B(m)$, $\forall m = 1, \dots, N$, where $\stackrel{d}{=}$ implies equality in a distributional sense.

Examples of reversible processes are linearly correlated stochastic processes and conservative chaotic systems; on the other hand, nonlinear stochastic processes and dissipative chaotic processes are generally irreversible. Moreover a thermodynamic interpretation [22] links the amount of irreversibility of a trajectory Σ to the amount of entropy that the underlying system is producing.

The typical approaches to evaluate time reversibility are based on a symbolization of the series and a statistical comparison of the symbol strings occurrence in the forward and backward series, or a compression algorithm [20]. The result depends on additional parameters such as the range partitioning or the size of the symbol alphabet; in particular long time series are required to estimate irreversibility in (discrete) stationary signals when the alphabet is large. Another issue that can arise is determined by the local nature of the symbolization process, as the presence of multiple scales could be swept away during the symbolization process (unless multi-scale algorithms are considered).

An alternative approach based on the visibility graphs and the Kullback-Leibler divergence has been proposed by Lacasa et al. [14] and will be discussed in more detail in section 3.2.1.

Chapter 3

Complex networks

3.1 Elements of graph theory

A *graph* $G = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ is defined as a set of *nodes* or vertices \mathcal{V} linked to each other by *edges* (or *links*) in the set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$; each edge is defined by the nodes it connects. In the case of a *directed* graph, an edge is defined as an ordered pair of nodes, referred to as outgoing and incoming node respectively; if the graph is instead *undirected*, the link is bilateral and defined by an unordered pair of nodes. An edge that joins a node to itself is called a *loop*. \mathcal{W} is a set of weights associated to the links; in the case of an *unweighted* graph \mathcal{W} coincides with the adjacency matrix, therefore each nonzero element is equal to 1 and represents an existing link. A *simple* graph is undirected, unweighted, with no loops and no *multiple edges*, meaning that for each pair of nodes there can be at most one edge.

A *walk* from node i to node j is a sequence $\gamma = (i = i_0, i_1, \dots, j = i_l)$ for which $(i_{h-1}, i_h) \in \mathcal{E} \quad \forall h = 1, \dots, l$; if such walk exists, node j is *reachable* from node i . A walk γ is a *path* from node i to j if it also holds that $i_h \neq i_k \quad \forall h, k \text{ s.t. } 0 \leq h < k \leq l$, with the possible exception of $i_0 = i_l$.

An undirected graph is *connected* if every node is reachable from any other node; similarly, a directed graph is *strongly connected* if given any pair of nodes i and j there is a path from i to j and vice versa.

Given a graph, an *induced subgraph* can be generated by selecting a subset of nodes and restricting the edge set accordingly; for a *spanning subgraph*, instead, it is sufficient to select a subset of the edges. It follows that if a monotonous property (for example being strongly connected) holds for a spanning subgraph $\tilde{G} \subseteq G$, then it also holds for G .

3.1.1 Measures and indices

A measure of node connectivity for an undirected graph is its *degree*, also known as degree centrality: it associates each node to the number of edges incident upon it. For a directed graph the degree of a node can be decomposed as the sum of its *out-degree* and *in-degree*, i.e. the number of outgoing and incoming edges from and to that node respectively.

To characterize the neighborhood of a node, one can refer to the *average nearest neighbors degree*:

$$\langle k_{nn} \rangle = \sum_{k'} k' P(k'|k), \quad (3.1)$$

where $P(k'|k)$ is the conditional probability that an edge of node with degree k points to a node with degree k' . Plotting this function over the degree can be useful to depict the overall assortativity trend for a network.

Alternatively one can estimate directly the *assortativity coefficient*, which is the Pearson correlation coefficient of degree between pairs of linked nodes:

$$r = \sum_{j,k} \frac{jk(e_{jk} - q_j q_k)}{\sigma_q^2}. \quad (3.2)$$

The term e_{jk} represents the fraction of edges that connect nodes of degree j and k , and q_k is the distribution of the *remaining degree*:

$$q_k = \frac{(k+1)p_{k+1}}{\sum_{j \geq 1} j \cdot p_j}, \quad (3.3)$$

where p_j is the probability for a node to have degree j .

For $-1 \leq r < 0$, the network is *disassortative*, meaning that low degree nodes are often connected with high degree nodes; $r = 0$ indicates non-assortativity, and for $0 < r \leq 1$ the network is *assortative*, indicating a stronger tendency for nodes to be linked with nodes of similar degree.

The concept of assortativity or assortative mixing is actually much broader and can be led by one or multiple discrete or scalar characteristic [19], but for the scope of this work it will be limited to assortative mixing by vertex degree.

3.2 Visibility Graphs

The *visibility graph* (VG) is a non-parametric algorithm first used by Lacasa et al.[1] to convert a scalar, univariate time series into a graph to describe its

structure and their underlying dynamics from a combinatorial perspective[8]. Every node is uniquely identified by a time series value y_i and its ordering index t_i , and the edges between nodes are determined according to a visibility criterion; for the natural visibility graphs it can be formalized as: given any two arbitrary nodes (t_a, y_a) and $(t_b, y_b) \in \{(t_i, y_i)\}_{i=1}^N$, they will have visibility if

$$y_c < y_b + (y_a - y_b) \frac{t_b - t_c}{t_b - t_a} \quad \forall (t_c, y_c) \mid t_a < t_c < t_b. \quad (3.4)$$

A simpler formulation of the VG algorithm is the *horizontal visibility graph* (HVG), proposed by Luque et al.; in this case a link between two nodes (t_a, y_a) and (t_b, y_b) exists if

$$y_c < \min\{y_a, y_b\} \quad \forall (t_c, y_c) \mid t_a < t_c < t_b. \quad (3.5)$$

To give an intuitive interpretation of the two visibility criteria, one can consider the nodes as a set of equispaced buildings along a straight line; the height of each building corresponds to the time series value¹. Given a direction of observation, an observer on the top of a building will be able to see only certain buildings: the first one in front of him will always be visible, and then he will have to keep raising his line of sight to look for buildings further away. In the case of a HVG, though, the observer can at most look straight ahead, so two buildings are mutually visible if and only if all the buildings in between are lower than them.

Following this idea, one can define an algorithm for the VG that for each node computes the slope of the line of sight between the first building and the following ones, and detects nodes with strictly increasing slopes to define the edges. On the other hand, the inspection of the following nodes for the HVG can be based directly on their node values. If the value of the fixed node is greater than that of the next node, it is sufficient to detect strictly increasing node values up to the first one greater or equal to the fixed node value; otherwise, only the first node after the fixed one is linked.

The algorithms presented for the construction of the VG and HVG should be interpreted as a naïve approach that can still handle a fairly long series - the longest tested has almost 21 thousand elements -, but it is not optimized for computational efficiency. A faster implementation would require a Divide&Conquer approach, as done by Iacobello [31],[32].

¹In this analogy one can assume positive values for the time series, but the same reasoning can be applied to negative values by setting the ground level to the lowest value of the time series

Algorithm 1: VG

Data: Array of node values $\{y_i\}_{i=1,\dots,N}$.

Result: 2D array *edges*, containing the end nodes indices for each link.

Initialize *edges*

for $k = 1 : N - 1$ **do**

$slopes = (y_i - y_k)/(i - k)$, with $i = k + 1 : N$

 Get $\{t_i\}_{i=1,\dots,n}$ corresponding to strictly increasing *slopes*_{*i*}

$t_i = t_i + k$

 Append to *edges* the rows (k, t_i) , with $i = 1, \dots, n$

Algorithm 2: HVG

Data: Array of node values $\{y_i\}_{i=1,\dots,N}$.

Result: 2D array *edges*, containing the end nodes indices for each link.

Initialize *edges*

for $k = 1 : N - 1$ **do**

if $y(k) > y(k + 1)$ **then**

 Find the first $kk > k$ so that $y_{kk} \geq y_k$ **if** $\nexists \quad kk$ **then**

$kk = N$

else

$kk = k + 1$

 Define $v = \{y_i\}_{i=k+1,\dots,kk}$

 Get $\{t_i\}_{i=1,\dots,n}$ corresponding to strictly increasing *v*_{*i*}

$t_i = t_i + k$

 Append to *edges* the rows (k, t_i) , with $i = 1, \dots, n$

else

 Append to *edges* the row $(k, k + 1)$

The graph extracted from a time series with a visibility method is always connected and undirected, as each node is connected to the following and visibility is defined as a mutual property between pairs of nodes. Even though there is no dependence on algorithmic parameters, there are potentially relevant boundary effects: the first point in the time series, for example, can only be visible to points in the future, limiting its degree.

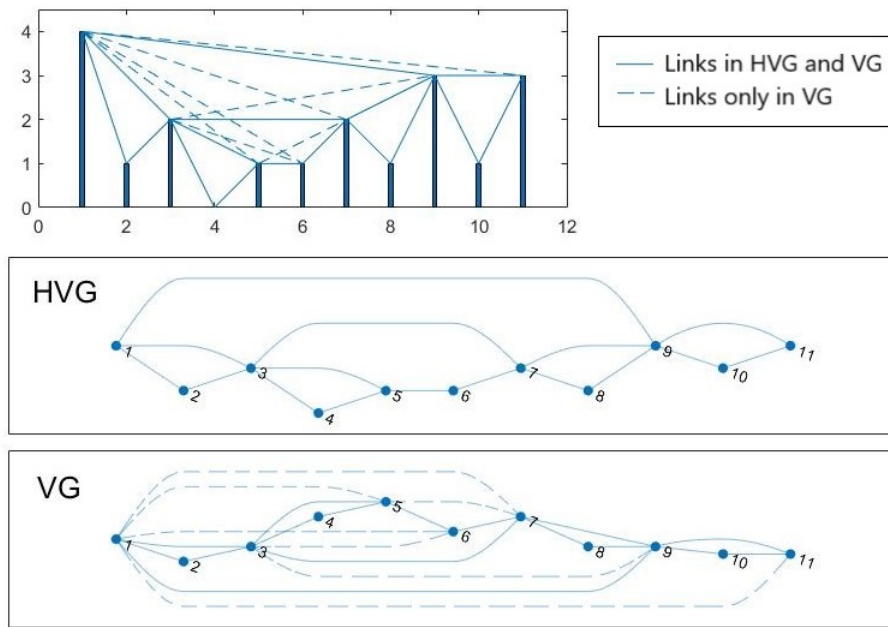


Figure 3.1: Example of a time series (11 data values) and the associated graphs derived from the horizontal visibility algorithm and the visibility algorithm. The visibility rays between the data define the links connecting nodes in the graphs.

Remark. For a given set of nodes, every link in the HVG is also a link in the VG but not vice versa; hence, the HVG can be considered as a spanning subgraph of the VG.

Proof. Assuming the same notation of (3.4) and (3.5), the statement is obvious for $y_c < y_b \leq y_a$.

For hypothesis (3.5) holds. Assume $y_c < y_a < y_b$; let $y_c = y_a - k$, with $k > 0$.

Add and subtract t_a from the fractional term in (3.4):

$$\frac{t_b - t_c}{t_b - t_a} = \frac{t_b - t_a + t_a - t_c}{t_b - t_a} = (1 - \epsilon), \quad \text{with } 0 < \epsilon < 1.$$

Substitute in (3.4):

$$y_a - k < y_b + (y_a - y_b)(1 - \epsilon) = y_b + y_a - y_b + (y_b - y_a)\epsilon.$$

Simplify the inequality and verify that it holds, as

$$-k < 0 < (y_b - y_a)\epsilon.$$

□

Natural visibility graphs are also invariant under affine transformations of the underlying time series, i.e. rescaling of both horizontal and vertical axes and horizontal and vertical translations; this property does not hold for horizontal visibility graphs though.

3.2.1 Irreversibility and Kullback-Leibler divergence

As already mentioned in section 2.3, relying on a symbolization of the time series to measure irreversibility may lead to some issues; hence it may be preferable to adopt a visibility graph approach instead. This requires the generation of the time directed counterpart to the (H)VG, in order to distinguish between incoming and outgoing links and calculate the in- and out-degree, also referred to by Donges et al.[6] as *retarded* and *advanced* degrees.

Kullback-Leibler divergence

To quantify the distinguishability between distributions, one can refer to the *Kullback-Leibler divergence* (KLD):

$$D[p(x)||q(x)] = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx, \quad (3.6)$$

where p and q are two generic distributions. $D = 0$ if and only if p and q are identical, otherwise it assumes positive values. The interpretation of the KLD as a measure of distinguishability is a consequence of the Chernoff-Stein lemma: the probability of incorrectly guessing (via hypothesis testing) that a sequence of n data is distributed according to p when the true distribution is q is asymptotically equal to $e^{-nD[p(x)||q(x)]}$. In addition, given X and Y random variables that describe the state of a system,

$$D[p(x, y)||q(x, y)] \geq D[p(x)||q(x)], \quad (3.7)$$

meaning that it is harder to distinguish between p and q when only marginal distributions are considered instead of the full joint distributions.

A method proposed by Lacasa et al.[14] estimates the irreversibility of the series by the Kullback-Leibler divergence (KLD) between the out- and in-degree distributions associated with the (H)VG, or by a generalized measure based on degree-degree distributions if needed.

Given the in- and out- degree distributions $P_{in}(k)$ and $P_{out}(k)$, the KLD measures the distance between them (in a distributional sense) as:

$$D[P_{in}(k)||P_{out}(k)] = \sum_k P_{in}(k) \log \frac{P_{in}(k)}{P_{out}(k)}. \quad (3.8)$$

In order to distinguish the degree of irreversibility, González et al.[8] introduce the *irreversibility ratio*² IR by standardizing the KLD with respect to a null model:

$$\text{IR} = \frac{\text{KLD}(in||out) - \langle \text{KLD}(in||out) \rangle_{null}}{\sigma[\text{KLD}(in||out)]_{null}}. \quad (3.9)$$

The terms $\langle \text{KLD}(in||out) \rangle_{null}$ and $\sigma[\text{KLD}(in||out)]_{null}$ are the mean and standard deviation of KLD of the null model, which is built by shuffling the time series to create a set of randomized samples. It is reversible by construction, hence its irreversibility value decreases as the time series length increases.

Interpreting the irreversibility ratio IR as a confidence index, a time series is (HVG) reversible if $\text{IR} \leq 1$, irreversible with *weak confidence* if $1 < \text{IR} \leq 4$, irreversible with *strong confidence* if $4 < \text{IR} \leq 10$ and irreversible with *extreme confidence* if $\text{IR} > 10$.

²The authors define more generally IR_m , but in this context only IR_1 is considered and the subscript is omitted.

Compared to other standard methods, a visibility graph approach to estimate time irreversibility requires a substantially smaller number of symbols, and can therefore be viable even with short time series [22]. The in- and out-degrees typically take values from a small alphabet, because the probability that an arbitrary node in a (H)VG has a certain in- and out-degree k typically decays exponentially fast with k [13].

This method is also robust to (reversible) noise pollution of the signal, unlike some standard approaches: even a small amount of noise can destroy the fractal structure of a chaotic attractor and mislead the calculation of chaos indicators such as the correlation dimension or the Lyapunov exponents[14].

3.2.2 Motifs and system dynamics

Motifs are small connected subgraphs consisting of a small fixed number of vertices (typically 3 or 4); ranking their relative frequency in descending order can provide information about the dynamic structure of the underlying time series. According to the so-called *superfamily phenomenon* of time series, different complex networks from the same type of flow data have the same rank ordering, and therefore belong to the same superfamily. Moreover within each superfamily networks corresponding to time series from different specific dynamic systems exhibit a unique fingerprint specific to that system [28].

There are 6 admissible motifs of size 4, presented in Fig.3.2. In particular motifs D and F lead the classification of the dynamics: the relative frequency of motif D increases from periodic to chaotic and finally to noisy periodic flows, whereas the relative frequency of the fully connected motif F decreases. The same trend is also observed passing from chaos to hyperchaos and then to noise, associated with an increase in total number of motifs detected.

In case of periodic flows the relative frequency of motifs D and E increases with the period; for periodic flows with increasing levels of noise, instead, the relative frequency of motifs C, E and F decreases whereas that of motif D increases. The only difference in case of correlated noise is a lower variability of relative frequency of motif E with the noise level.

Remark. *A HVG cannot have a fully connected 4-nodes subgraph; hence, motif F is impossible.*

Proof. Let us assume that a four-nodes fully connected subgraph of a HVG

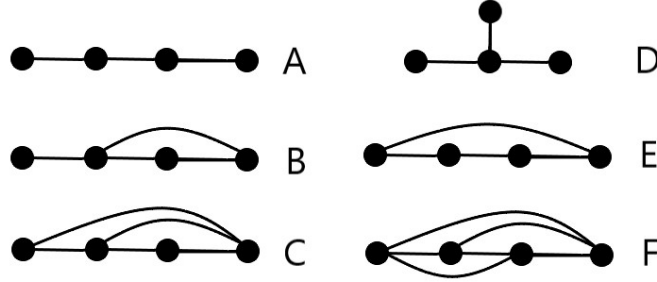


Figure 3.2: Classification of 4 node motifs for undirected graphs. Each one is labelled with a letter from A to F.

exists; and let $\{y_i\}_{i=1}^4$ the node values, taken at time t_i . For the definition of fully connected graph, all pairs of nodes are linked, and each edge satisfies the HVG linking criteria (3.5).

The edge between node 1 and 3 implies $y_2 < \min\{y_1, y_3\}$, and in particular $y_2 < y_3$; similarly, the edge between node 2 and 4 implies $y_3 < \min\{y_2, y_4\}$, and so $y_3 < y_2$, which is absurd. \square

Sequential motifs

The sequential n -node motifs are characterized by node labels appearing in strict sequential order, and can be detected by checking iteratively the links between nodes selected by a sliding a window of size n . In this way the dynamical information of the series is preserved and the computation runs in linear time $O(N)$, where N is the size of the time series. In particular for the HVGs the motif classification can be performed by analyzing the time series values through a set of inequalities presented in Fig.3.3; there are 2 admissible motifs of size 3, and 6 admissible motifs of size 4.

The *motif significance profile*, also known as simply motif profile, is defined as the vector function $\mathbf{Z}^n : n \in \mathbb{N} \rightarrow [\mathbb{P}_1^n, \dots, \mathbb{P}_p^n] \in [0, 1]^p$ that associates the motif size n to the relative frequency \mathbb{P}_i^n of each type- i motif. The HVG motifs induce a particular partition of the set of ordinal patterns, and the analysis of the motif profiles can distinguish between different types of complex dynamics. Type-II 4-node motif, for example, is absent for irregular (aperiodic) real-valued time series. It can be proven in the limit of infinite-size series [10] that i.i.d. - e.g. Gaussian, uniform, power law, etc., uncorrelated random series- all have the same HVG motif profiles:

$$\mathbf{Z}^3 = \left[\frac{2}{3}, \frac{1}{3} \right]; \quad \mathbf{Z}^4 = \left[\frac{8}{24}, 0, \frac{6}{24}, \frac{6}{24}, \frac{2}{24}, \frac{2}{24} \right]. \quad (3.10)$$









| Motif label | Motif type | Inequality set |
|-------------|---|--|
| 1 |  | $\{\forall(x_0, x_2), x_1 > x_0\} \cup \{\forall x_0, x_1 < x_0, x_2 < x_1\}$ |
| 2 |  | $\{\forall x_0, x_1 < x_0, x_2 > x_1\}$ |
| 1 |  | $\{\forall(x_0, x_1), x_2 < x_1, x_3 < x_2\} \cup \{\forall(x_0, x_3), x_1 > x_0, x_2 > x_1\}$ |
| 2 |  | $\{\forall x_0, x_1 < x_0, x_2 = x_1, x_3 > x_2\}$ |
| 3 |  | $\{\forall x_0, x_1 < x_0, x_1 < x_2 < x_0, x_3 < x_2\} \cup \{\forall(x_0, x_3), x_1 < x_0, x_2 > x_0\}$ |
| 4 |  | $\{\forall x_0, x_1 > x_0, x_2 < x_1, x_3 > x_2\} \cup \{\forall x_0, x_1 < x_0, x_2 < x_1, x_2 < x_3 < x_1\}$ |
| 5 |  | $\{\forall x_0, x_1 < x_0, x_1 < x_2 < x_0, x_3 > x_2\}$ |
| 6 |  | $\{\forall x_0, x_1 < x_0, x_2 < x_1, x_3 > x_1\}$ |

Figure 3.3: Enumeration of all sequential 3- and 4-node motifs for the visibility graphs. Each motif can be characterized according to a hierarchy of inequalities in the associated time series [10]. The values $\{x_i\}_{i=0}^3$ refer to 4 consecutive values in the time series.

Similarly, the motif profile for the fully chaotic logistic map - as an example of deterministic chaos - is:

$$\mathbf{Z}^3 = \left[\frac{2}{3}, \frac{1}{3} \right]; \quad \mathbf{Z}^4 = \left[\frac{8}{24}, 0, \frac{4}{24}, \frac{8}{24}, \frac{4}{24}, 0 \right], \quad (3.11)$$

and it can be noted how the two processes can be distinguished by comparing the last 4 components of the 4-node motif profiles.

Lastly, an example of stochastic process with correlation is considered; given Gaussian white noise $\xi_t \sim \mathcal{N}(0, 1)$ and a correlation parameter $r \in (0, 1)$, the colored noise with exponentially decaying correlations is described by the AR(1) process:

$$\begin{aligned} x_0 &= \xi_0, \\ x_t &= r x_{t-1} + \sqrt{(1 - r^2)} \xi_t, \quad t \geq 1. \end{aligned} \quad (3.12)$$

For $r \rightarrow 0$ the process tends to a white noise signal, and for $r \rightarrow 1$ the process gets completely correlated and tends to be constant $x_{t+1} = x_t \quad \forall t$, as shown by the 4-node motif profiles reported in Fig.3.4.

This method of dynamics discrimination based on motif profiles can be used to analyze empirical time series as convergence to the asymptotic theory is already reached for series of size $N \ll 10^4$, and the discrimination between dynamics is robust to measurement noise pollution. For example, given a chaotic signal $x_t = 4x_{t-1}(1 - x_{t-1})$ and a uniform white noise signal $\xi \sim U[0, a]$, $0 \leq a \leq 1$, a noisy chaotic signal $Y(t) = x_t + \xi$ can be correctly

| r | \mathbb{P}_1^4 | \mathbb{P}_2^4 | $\mathbb{P}_3^4, \mathbb{P}_4^4$ | $\mathbb{P}_5^4, \mathbb{P}_6^4$ |
|------|------------------|------------------|----------------------------------|----------------------------------|
| 0.02 | 0.3370 | 0 | 0.2482 | 0.0833 |
| 0.04 | 0.3406 | 0 | 0.2464 | 0.0833 |
| 0.06 | 0.3443 | 0 | 0.2446 | 0.0832 |
| 0.08 | 0.3478 | 0 | 0.2429 | 0.0831 |
| 0.1 | 0.3514 | 0 | 0.2412 | 0.0830 |
| 0.2 | 0.3690 | 0 | 0.2333 | 0.0822 |
| 0.3 | 0.3862 | 0 | 0.2260 | 0.0809 |
| 0.4 | 0.4030 | 0 | 0.2192 | 0.0793 |
| 0.5 | 0.4196 | 0 | 0.2130 | 0.0772 |
| 0.6 | 0.4359 | 0 | 0.2072 | 0.0748 |
| 0.7 | 0.4521 | 0 | 0.2018 | 0.0722 |
| 0.8 | 0.4681 | 0 | 0.1967 | 0.0692 |
| 0.9 | 0.4841 | 0 | 0.1920 | 0.0660 |
| 0.95 | 0.4919 | 0 | 0.1897 | 0.0643 |
| 0.97 | 0.4945 | 0 | 0.1888 | 0.0636 |
| 0.99 | 0.4973 | 0 | 0.1879 | 0.1879 |

Figure 3.4: Theoretical values of $\mathbf{Z}^4(r)$ for the AR(1) process evaluated at different values of the coefficient r .

classified for a noise-to-signal ratio³ $NSR \approx 2.67$.

³ $NSR = \sigma_\xi^2 / \sigma_Y^2$, where σ_i^2 is the variance of signal i .

Chapter 4

Case study 1: Turin 1753-2020

4.1 Series description

The series have been retrieved and elaborated by Di Napoli and Mercalli in the context of a long term study of the climate of Turin, and analyzed in great detail in the book *Il clima di Torino* [17]. They consist of daily maximum and minimum temperature measurements, from the year 1753 up to 2020. Each observation is paired with the location in which it was performed; as shown in Fig. 4.1, the majority of data is collected in the city of Turin, but there have been a few station relocations over time. The interval from 1753 to 1786 is the least homogeneous, with frequent changes in location over a vaster area in Piedmont; this is due to the fact that the measurements were performed by Ignazio Somis, who as the king's physician had to frequently follow him all around Piedmont. Nevertheless, the fact that many stations share similar conditions, such as elevation above the ground and location in an urban area, results in an impact on the homogeneity that is smaller than expected. The series is also characterized by inhomogeneities in the measurement procedures: the maximum-minimum thermometer was introduced only in 1857, and the unit used until 1848 was Réaumur degrees as an alcohol thermometer was used. Moreover both the maximum and minimum temperature records present missing values in the period 1753-1865, as reported in table 4.1.

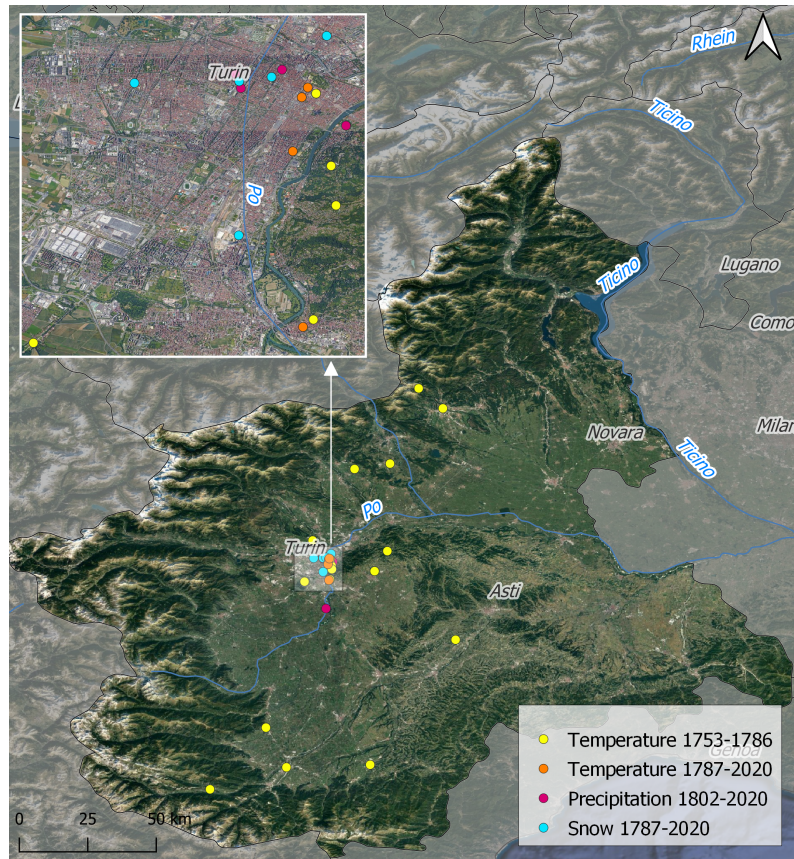


Figure 4.1: Locations for the temperature, precipitation and snow observations of the series Turin 1753-2020.

| Variable | Miss. values | Longest missing interval | Duration (days) |
|----------|--------------|---------------------------|-----------------|
| minT | 783 | 13 Jun 1775 - 22 Oct 1775 | 131 |
| maxT | 871 | 12 Jun 1775 - 22 Oct 1775 | 132 |

Table 4.1: Missing value details for the temperature series

4.1.1 Homogenization

The temperature series is presented in both its “original” and homogenized version. The “original” series is the result of:

- correction/reduction of errors of registration, annotation and publication of the daily values;
- conversion to °C;
- restoration where possible of a 24 hour observation period for the measurement of daily extremes (concerning in particular the period 1885-1961).

The homogenized series has been virtually lead back to the station of the Ufficio Idrografico del Po (UIPO) in corso Bolzano, where the observations from March 1961 to December 2004 were conducted, and at the current urban expansion of the city. The following methods have been applied to calculate corrections to the “original” series:

- gap completion, estimating the missing days from nearby records;
- discontinuity detection and reduction, addressing each one of the following causes separately:
 - a) missing record of daily minimum temperature before 10.Feb.1857 and maximum temperature before 01.Aug.1857;
 - b) thermometer replacement/relocation, change of exposition, etc.;
 - c) change in stations;
 - d) urban expansion, in particular construction of new buildings.

The homogenization has proven challenging for data up to 1866, as there are very few series available for reference, and they are affected as well by discontinuities. Therefore, the estimation of gaps for this first leg of the series has been performed with the aim of simply returning homogeneity on the daily thermal excursion in the intervals with annual average way above or

below the norm of the corresponding station. The rest of the data, instead, can be compared with multiple series from stations in the Po basin, instituted after the establishment in 1865 of the national meteorological service. The comparison is performed with a Craddock test, and the correction is computed as a constant to sum on the non-homogeneous interval detected. The daily temperature extremes for the first century of the series were estimated from two or three daily observations at fixed hours, depending on the normal daily thermal oscillation typical of that period of the year and on the atmospheric conditions of the single day. This implies that some anomalies in the daily temperature have not been recorded, and as such some extremes may be incorrectly estimated; nevertheless, given the relative infrequency of anomalies in the climate of Turin, the overestimation of minimum temperature and underestimation of maximum temperature on the annual average is considered negligible ($\leq 0.1^\circ\text{C}$).

Change in stations

To correct the inhomogeneities introduced by a change in stations, a comparison between each station in closing phase and the following one has been performed; in case of no simultaneous observations, a third station has been considered. The difference of maximum (minimum) temperature between the two stations is calculated, and the days with a value below the 5th percentile or above the 95th percentile¹ of the monthly series are eliminated. For the remaining days, the average temperature $\bar{T} = (T_{max} + T_{min})/2$ and daily thermal excursion $DTE = T_{max} - T_{min}$ are calculated for each station.

A least squares approximation is used to estimate the regression function of average temperature between simultaneous observations in the two stations over each month:

$$\hat{\hat{T}}_y = a + b \cdot \bar{T}_x, \quad (4.1)$$

where the subscripts x and y refer to the old and new station respectively. The bilateral comparison is then performed between $\hat{\hat{T}}_y$ and \bar{T}_y to adjust the parameters of the regression lines. The results combined provide for each station different from UIPO a set of 12 regression functions of type (4.1) that perform for each day the virtual change of stations to the reference one; possible discontinuities in correction at the change of the month have

¹The thresholds are set at the 10th and 90th percentile in case of stations particularly far or with big differences in altitude.

been corrected with a moving average smoothing with a 31-day window. The daily thermal excursion estimation for the new station is estimated as

$$\widehat{DTE}_y = RDTE \cdot DTE_x, \quad (4.2)$$

where $RDTE$ is the monthly average of the daily thermal excursion ratio DTE_y/DTE_x smoothed over the year with a moving average with a 61-day window. The smoothed $RDTE$ coefficients computed from the bilateral comparisons are multiplied to obtain for each station a $RDTE$ for each day of the year that allows to convert the DTE values so that they can be assigned to the UIPO station.

The series of daily average temperature and daily thermal excursion estimated for the UIPO station allow for the estimation of maximum and minimum temperature:

$$\begin{aligned} T_{max} &= \hat{T} + \widehat{DTE}^2, \\ T_{min} &= \hat{T} - \widehat{DTE}^2; \end{aligned} \quad (4.3)$$

Urban expansion

The main effect of urban expansion on temperatures is the *urban heat island* (UHI), i.e. a difference in temperature - always positive difference in the annual average - between the urban conglomerate and the surrounding rural area, caused by a larger house density, industrial activity and traffic emissions. During the period 1990-2002, for example, the rural areas surrounding Turin was about 1.9°C colder than the city. This inhomogeneity factor has already removed for measurements recorded after 1960, so only the period 1753-1960 is considered. The maximum (minimum) temperature is referred to the current level of urbanization by adding the difference between the current urban heat island intensity and the one present at the time of measurement²:

$$\hat{T}_{current} = T_{old} + (UHI_{current} - UHI_{old}). \quad (4.4)$$

The urban heat island intensity can be estimated from the number of urban residents P [5] with the empirical formula:

$$UHI = \alpha \cdot \log P, \quad (4.5)$$

²The heat island varies too much on a daily basis, so a monthly average level is considered.

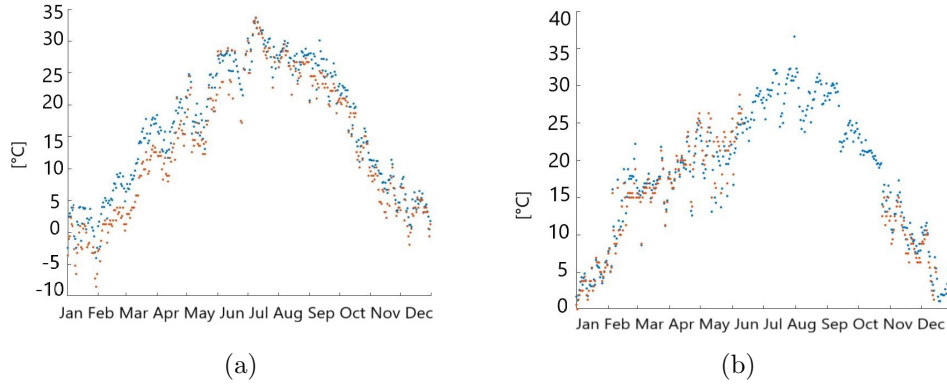


Figure 4.2: Original (orange) and homogenized (blue) series of daily maximum temperature for the year 1753 (a) and 1775 (b).

where α is a parameter specific to the city. More precisely, the heat island is specific of each urban-rural pair of stations, as proven by the weak correlation identified on a study of pairs of urban-rural stations in Novara, Milano, Brescia, Bologna and Parma [Zanella 1976, Grillini 1978, Bottau 1997, Beltrano & Perini 1996].

The homogenization of the temperature series results on average in a positive correction for the maximum series and a negative one for the minimum; as expected, the difference between the homogenized and original series is greater in the first years, due to the presence of gaps in the data and higher variety in the spatial distribution of the observation sites. The main statistics of the difference between homogenized and original series are reported in tables 4.1.1 and 4.1.1, and an example of comparison between the original and homogenized maximum temperature series is presented in Fig.4.2; in particular Fig.4.2b shows how the homogenized series fills the longest missing data interval in the series.

| Homogenized-Original maxT | | |
|---------------------------|-----------|-----------|
| Metric | 1753-1865 | 1866-2020 |
| Mean | 0.9246 | 0.2950 |
| Std. dev | 1.411 | 0.8879 |
| Max | 8.800 | 4.400 |
| Min | -7.800 | -2.600 |
| 90th prctile | 2.700 | 1.500 |

| Homogenized-Original minT | | |
|---------------------------|-----------|-----------|
| Metric | 1753-1815 | 1816-2020 |
| Mean | -0.8284 | -0.1990 |
| Std. dev | 1.529 | 1.064 |
| Max | 4.100 | 4.600 |
| Min | -12.50 | -9.500 |
| 90th prctile | 0.9000 | 0.9000 |

Table 4.2: Mean, standard deviation, maximum, minimum and 90th percentile of the difference between homogenized and original series of maximum temperature (above) and minimum temperature (below). The division in two periods is defined in order to have missing data only in the first time interval.

4.2 Preliminary analysis

As the scope of this work revolves around the use of networking techniques to extract information about the underlying time series, the following analysis will focus exclusively on the homogenized version of the maximum and minimum temperatures; for simplicity, the adjective homogenized is henceforth omitted.

The maximum temperature ranges from -7.8°C to 39.7°C , recorded on January 25 1758 and August 11 2003 respectively. The minimum temperature instead ranges from -21.2°C to 26.8°C , recorded on February 3 1754 and July 7 2015 respectively. Both series can be fitted with a bimodal distribution, as shown in Fig.4.3; the peaks for the maximum temperature distribution are 9 and 26°C , and 8 and 16°C for the minimum temperature.

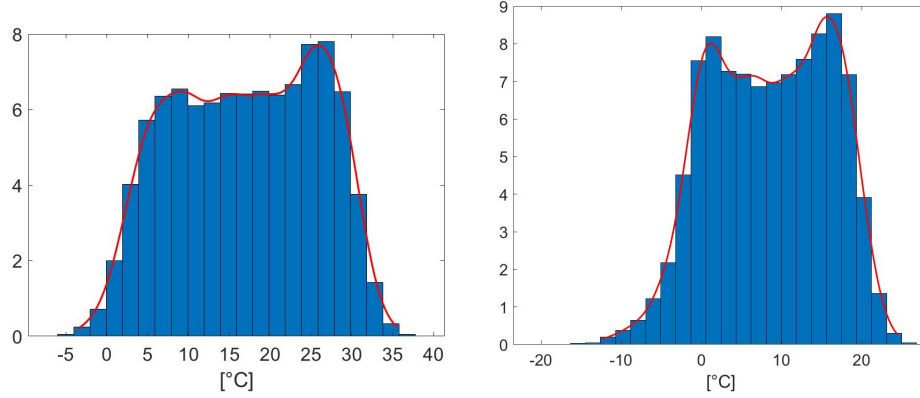


Figure 4.3: Relative frequency (%) of daily maximum (left) and minimum (right) temperatures. The red curve is the bimodal distribution fitted to the histogram.

The monthly average of maximum and minimum temperatures are lowest in January and highest in July; the results, divided between values before and after 1901 are presented in Fig.4.4. The monthly averages are lower for the most recent period, with the exception of September and October for the maximum temperature. The corresponding monthly standard deviation plots of maximum and minimum temperature are presented in Fig.4.5; the standard deviation of maximum temperature in the period 1901-2020 is lower than the corresponding value for the period 1753-1900. The same applies to the plot relative to the minimum temperatures, for months May to September.

The yearly record of maximum temperature falls between June and August, with 50% of the instances concentrated in July and an exception of 4 occurrences in May and 4 in September over a period of 268 years. Analogously, most yearly records of minimum temperature occur between December and February, with 55% in January; only 4 years register their coldest date in November and 6 in March.

The yearly average minimum temperature ranges from 6.6°C, recorded in 1855, to 11.1°C in 2020; the yearly average of maximum temperature, instead ranges from 14.76°C, recorded in 1814 and 1855, to 19.65°C, recorded in 2015. As shown in Fig.4.6a, the 9-year moving average of the minimum temperature shows periodic oscillations between approximately 7°C and 9°C; in the last century, though, the amplitude of the oscillations is

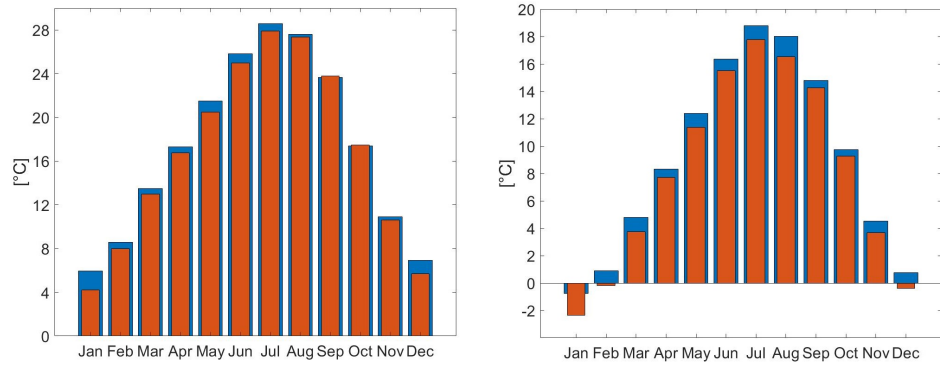


Figure 4.4: Monthly mean of maximum (left) and minimum (right) temperatures. The blue series refers to the period 1753-1900, and the orange one to 1901-2020.

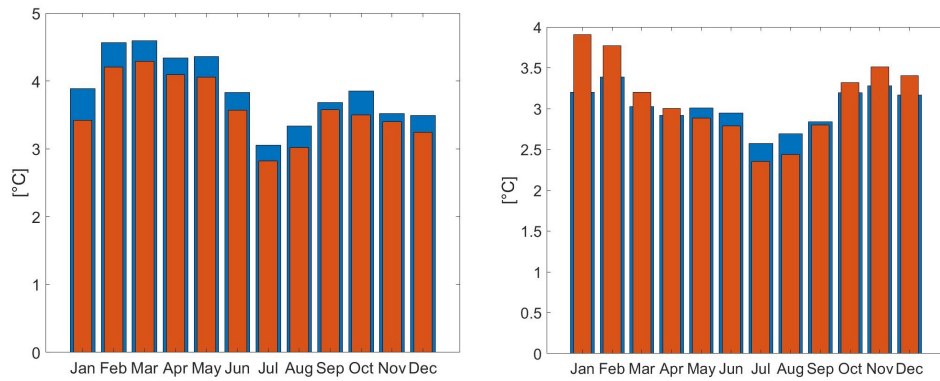
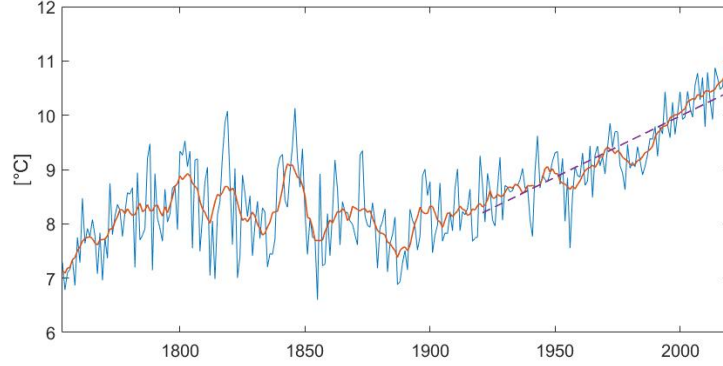
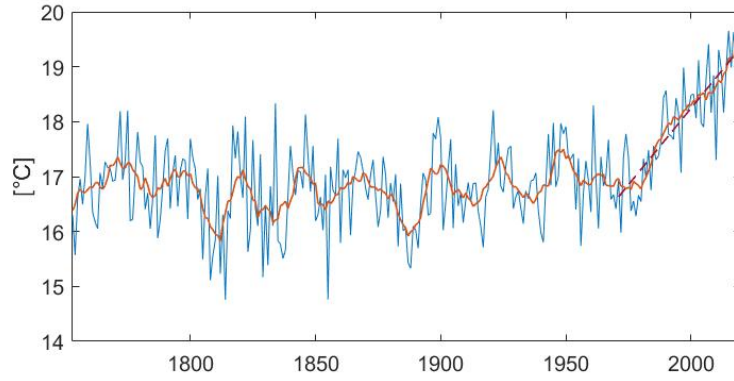


Figure 4.5: Standard deviation of monthly maximum (left) and minimum (right) temperatures. The blue series refers to the period 1753-1900, and the orange one to 1901-2020.



(a) Yearly average of minimum temperature (blue line), 9-year moving average (red line) and linear trend for the last 100 years of the series (red dashed line).



(b) Yearly average of maximum temperature (blue line), 9-year moving average (red line) and linear trend for the last 50 years of the series (red dashed line).

reduced and the smoothed data follows a positive linear trend³. Similarly, the yearly average maximum temperature and its 9-year moving average is presented in Fig.4.6b: the moving average shows periodic oscillations between 16°C and 17.5°C, with the exception of the last 50 years; in this period the smoothed data follows a linear trend that goes from 16.63°C to 19.63°C.

As a result of the increasing temperatures, the number of frost days in a year has significantly decreased over time especially in the last century, as shown in Fig.4.7. The highest number of days with maximum temperatures

³All linear trends are tested at the 1% significance level.

below 0°C is 35 in 1755, and before 1950 several peaks with values between 14 and 25 frost days are recorded; the last peak at 14 days occurs in 1956, and the following local maxima decrease in value, reaching 10 days in 2010. The number of days in a year with maximum temperature over 30°C has increased significantly⁴ over the last 30 years, as shown in Fig.4.8. The number of hot days is quite high at first, reaching a peak of 52 days in 1772, and it decreases subsequently, until new peaks of 44 and 42 days in 1928 and 1945 respectively. Since 1985 summers have become increasingly longer, with 1991 almost reaching the previous record of 1772 and then 2003 featuring a record of 72 hot days, followed by other 6 years recording at least 52 hot days in summer.

A similar trend is reflected in the analysis of heat waves according to the definition of Mercalli[17], as shown in Fig.4.9. The maximum intensity of heat waves first assumes values relatively high, with a peak of 36.35°C in 1771, then decreases in the following century, and increases again in the last 150 years, with a new record of 37.57°C set in 2003 and intensities never lower than 32°C from the year 2000. The year 2003 also features a peak of heat wave duration of 37 days, previously matched only in 1928, and later surpassed in 2010 by a new record of 47 consecutive days with daily maximum temperatures greater or equal to 28°C . The instances of years with no heat wave detected all belong to the periods 1806-1883 and 1908-1977, with most concentrated in the past.

Volcanic eruptions

The average temperature series is analysed by Mercalli [17] also in relationship with the major volcanic eruptions of the last two centuries; the most notable effects are observed in correspondence of the eruption of the Tambora in 1815 and of the Coseguina in 1835. Turin did not experience the “year without summer” in 1816, as the temperatures remained quite mild, but an exceptional drought occurred from August 1816 to February 1818, registering an amount of precipitation equal to 43% of the average of the period 1803-2007. On the other hand, an intense cooling occurred between 1835 and 1838: the yearly average temperature during that period never went above 11.6°C , and the coldest winter of the entire series occurred in 1835.

⁴with a confidence level $> 99\%$ according to the Mann-Kendall test

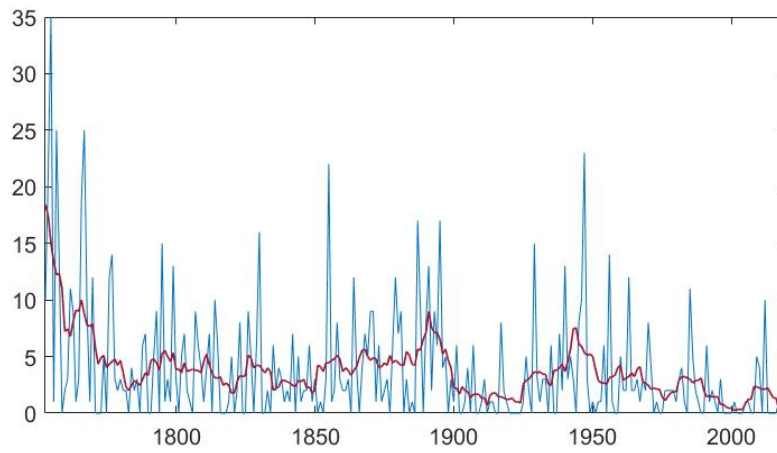


Figure 4.7: Number of frost days in a year (blue line) and smoothed average with a 9-year window (red line).

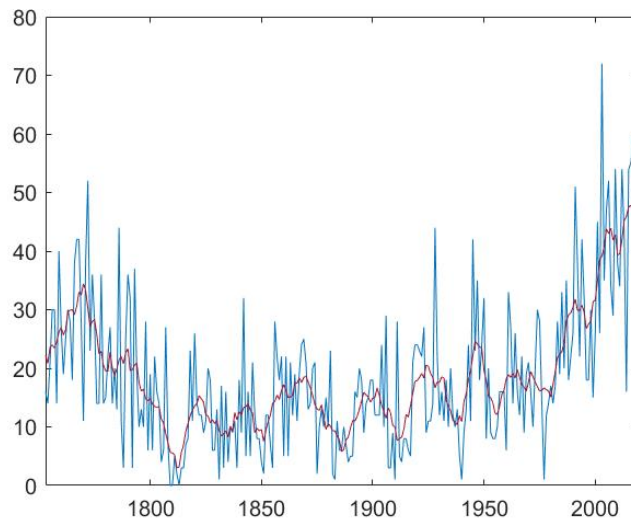


Figure 4.8: Number of days in a year with daily maximum temperature above 30°C (blue line) and smoothed average with a 9-year window (red line).

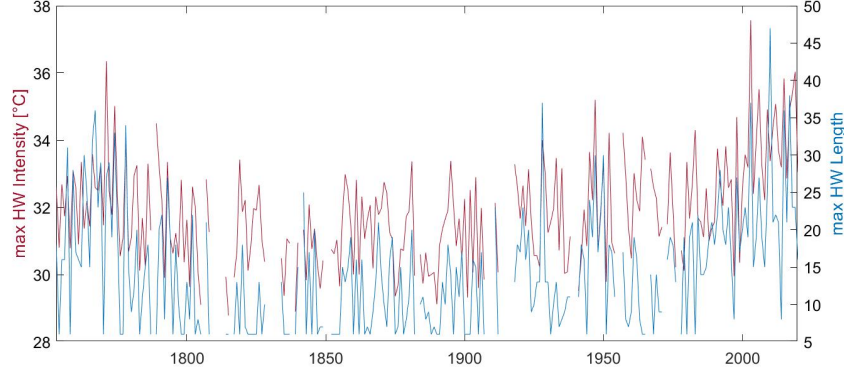


Figure 4.9: Maximum intensity of the heat waves detected each year, as defined by Mercalli [17], and maximum length of the heat wave, i.e. number of consecutive days with maximum daily temperatures $\geq 28^\circ$.

4.3 Visibility graphs

The natural and horizontal visibility graphs considered for the analysis of the maximum temperature series are computed over the full series and over a partition of the data in 5 and 10 intervals⁵, meaning that for each interval a graph is generated taking a subset of the time series as nodes. The full series and its partitions are then detrended to generate two detrended versions of each graph; for the first version the trend is computed on the whole series, whereas for the second version (indicated as “DT loc”) each interval of the partitions is detrended separately. Moreover for each base graph a deseasoned version is generated: the series or its partition is deseasoned by computing the “average year”, i.e. the mean of temperatures for each day of the year over the given period, and subtracting from each data point the corresponding average temperature.

4.3.1 Degree metrics

The first four moments of the degree for the HVG, VG and their detrended and deseasoned counterparts computed on the full series are presented in

⁵It is the same division used in table ??, and the 5 interval set is derived by merging pairs of intervals together. With an abuse of language these graphs will also be identified by referring to the interval of the underlying time series.

table 4.3. All the metrics for the VGs are greater than the corresponding value for the HVGs. There is not much difference in terms of moments between the different versions of HVG; each moment for the deseasoned HVG is slightly greater than the corresponding value for the detrended HVG, and these are slightly greater than the corresponding value for the HVG computed on the original series. Also the moments of the degree of the VG and its detrended counterpart are close in value, but there is a greater difference w.r.t. the deseasoned VG. More specifically, the degree mean and standard deviation of the deseasoned VG are lower w.r.t. the VG, whereas the skewness and kurtosis are greater. This suggests that the degree distribution for the deseasoned VG has a lower peak, is more skewed to the right and has more outliers, corresponding to nodes in the graph with degree much greater than the average.

The same metrics are shown in figure 4.10 for the graphs computed on 5 or 10 intervals. Each metric is characterized by a significant linear trend, that is positive for the mean of the HVGs, the skewness and kurtosis of the VGs, and negative in all other cases. The linear trend for the mean degree of the detrended and deseasoned HVG feature the smallest slope in absolute value. On the other hand, the kurtosis plots feature trends with the greatest slopes in absolute value, with the exception of the deseasoned VG. The deseasoned graphs always feature a weaker trend in comparison to their original counterpart. The choice between 5 or 10 intervals does not particularly affect the results for the HVGs, but for the VGs the linear fit of skewness and kurtosis improves when longer intervals are considered. The global and local detrending of the series lead to almost identical results in terms of moments of the degree of the corresponding graphs.

Assortativity

The neighborhood connectivity of the HVGs and VGs computed over 5 and 10 intervals are reported in figures 4.11 and 4.12 respectively. The partition of the time series in 5 or 10 intervals does not particularly affect the plots, with the exception of a few outliers for the highest degree values in the deseasoned HVG and the VG; also the differences between detrending methods are minimal. The plots for the HVGs have a much lower number of data points compared to those for the VGs, implying that the number of unique degree values is a lot lower; moreover the maximum degree ranges from 23 to 30 for the HVGs, and from 131 to 154 for the VGs. This is in accord with the notion that the HVG can be interpreted as a subgraph of

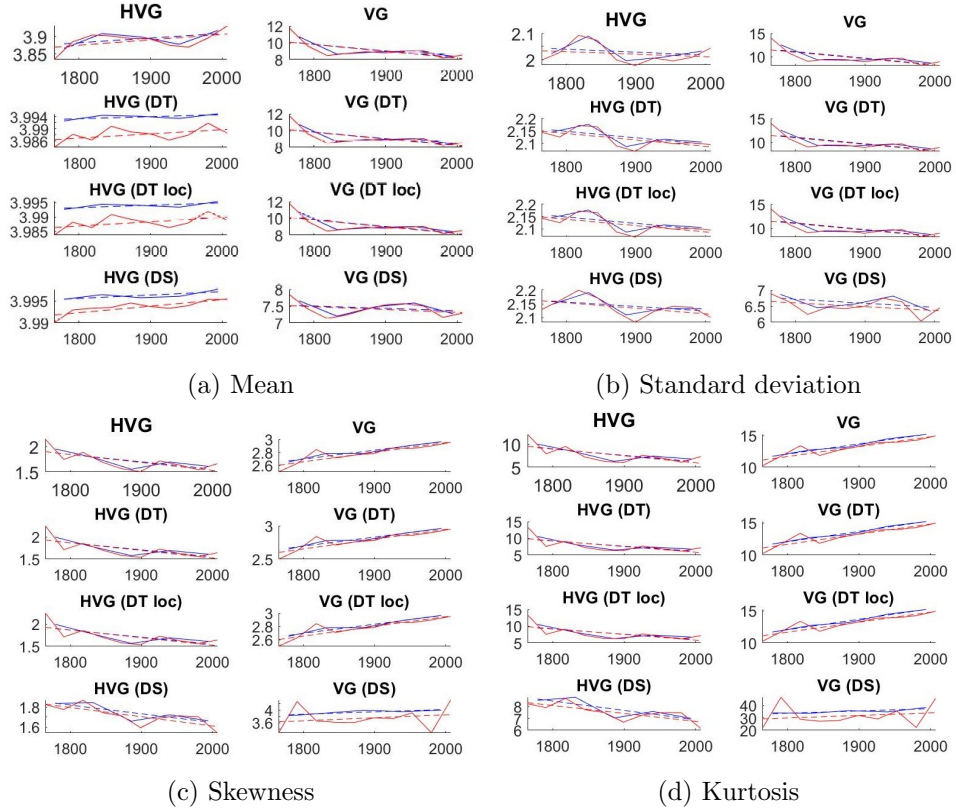


Figure 4.10: First four moments of the degree for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 5 intervals (blue line) and 10 intervals (red line). The corresponding linear trends are highlighted with a dashed line when significant.

Degree Metrics

| Graph | Mean | Std. deviation | Skewness | Kurtosis |
|----------|-------|----------------|----------|----------|
| HVG | 3.898 | 2.035 | 1.724 | 7.803 |
| HVG (DT) | 3.999 | 2.133 | 1.738 | 7.926 |
| HVG (DS) | 3.999 | 2.147 | 1.766 | 7.996 |
| VG | 9.163 | 10.09 | 2.920 | 14.65 |
| VG (DT) | 9.168 | 10.09 | 2.918 | 14.62 |
| VG (DS) | 7.485 | 6.716 | 4.070 | 37.92 |

Table 4.3: First four moments of the degree for the HVG, VG and their detrended (DT) and deseasoned (DS) counterparts, computed over the full series.

the VG, where only a subset of the edges is kept.

The neighborhood connectivity plots for the HVGs display points above the bisector only for degree values lower than 5; similarly, the plot points for the deseasoned VG are above the bisector only for degree values below 13. For the VG, instead, the bisector is crossed for degree values between 21 and 31 depending on the time interval considered; the most recent series intersect the bisector for lower degree values, and then the crossing threshold is gradually shifted towards higher degrees. The same observation applies to the detrended VGs, with the interception of the bisector occurring between degree 22 and 30.

The NC plots for the HVGs are distributed along a line that is almost horizontal for the HVG and its detrended counterparts and has positive slope for the deseasoned one; the points are most scattered for the least recent intervals, especially for degree values between 17 and 22. The plots for the VG and detrended VG feature a larger variability in NC between different time intervals and within the same series from values of degree between 40 and 60.

Interestingly, the oldest interval of both partitions (1753-1804 and 1753-1777) stands out in the plots for the VG and its detrended counterparts: not only do the major outliers belong to this series, but the series as a whole lies almost always above all the other data points, i.e. for any given degree the oldest nodes tend to have the highest neighbourhood connectivity. In the case of the deseasoned VG, instead, all series follow a positive linear trend, with some contained variability between degree 40 and 90; this is the instance with highest disassortativity for high degree nodes.

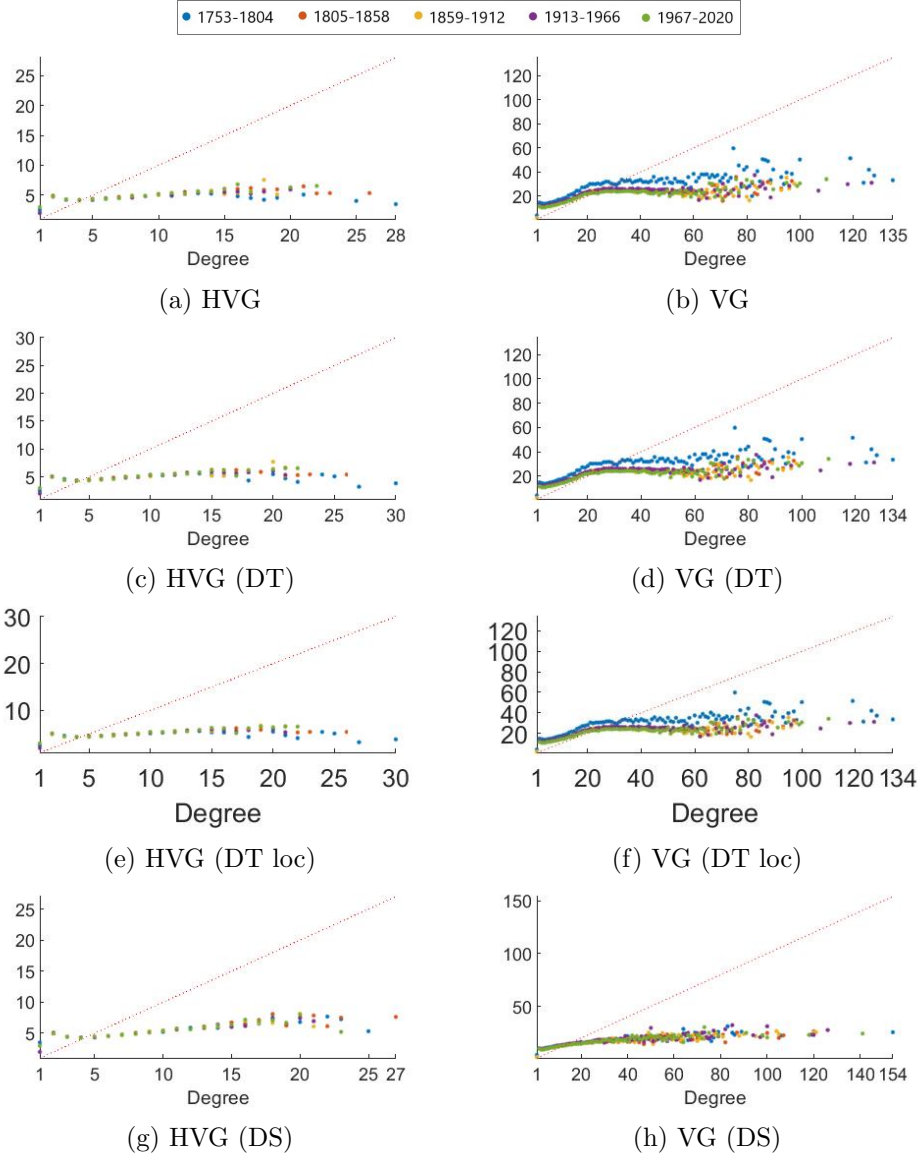


Figure 4.11: Neighborhood connectivity for the HVG, VG and their de-trended (DT/DT loc) and deseasoned (DS) counterparts, computed over 5 intervals. The bisector $y = x$ is highlighted with a red dashed line for each plot.

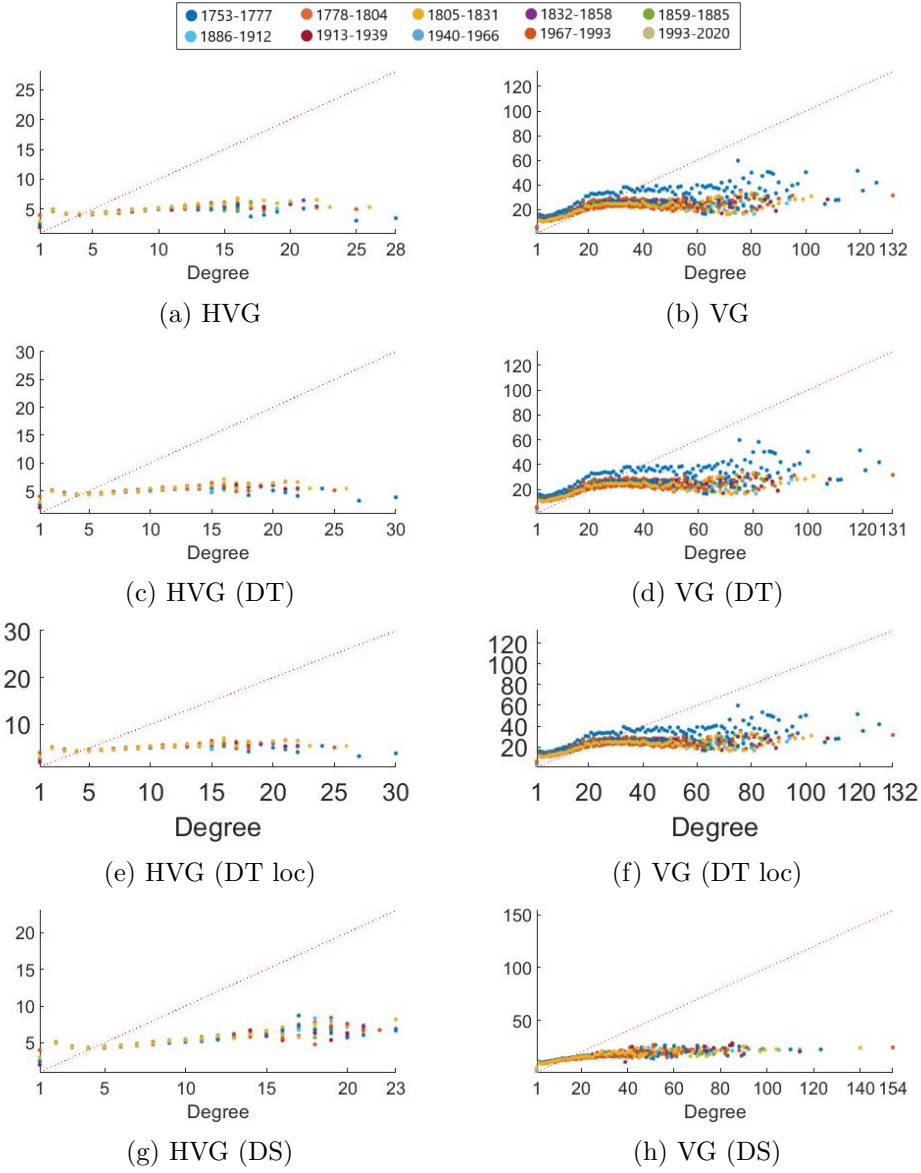


Figure 4.12: Neighborhood connectivity for the HVG, VG and their de-trended (DT/DT loc) and deseasoned (DS) counterparts, computed over 10 intervals. The bisector $y = x$ is highlighted with a red dashed line for each plot.

Time reversibility

The time reversibility test is performed by calculating the Kullback-Leibler divergence based on the in- and out-degree of the HVG, VG and their detrended and deseasoned counterparts; these values are standardized w.r.t. a null model obtained by shuffling the corresponding series, and the resulting irreversibility ratios IR are presented in table 4.5.

The test on the VG returns IR values associated with irreversibility with strong and extreme confidence for both partitions of the series considered. The HVG instead features an interval (1859-1912) that is reversible with weak confidence and corresponds in the finer partition to two reversible intervals; moreover in this latter case also the period from 1967 to 2020 is reversible with weak confidence. Similarly, the detrended VG in the 5-interval partition is completely irreversible, and in the 10-interval partition the periods 1859-1912 and 1967-1993 are weakly reversible. The detrended HVG is reversible with at least weak confidence in the periods 1805-1858, 1913-2020; in particular the subintervals 1805-1831, 1940-1966 and 1967-1993 are reversible. The deseasoned HVG is reversible on the period 1859-1912 for both partitions considered; in the finer partition also the period 1967-2020 is reversible with at least weak confidence. The deseasoned VG in the 5-interval partition is reversible with at least weak confidence from 1805 to 2020, which is reflected in the finer partition, with the exception of the interval 1940-1966.

Irreversibility ratio (IR)

| Period | HVG | HVG_DS | HVG_DT loc | HVG_DT |
|-----------|---------|--------|------------|---------|
| 1753-1777 | 22.06 | 23.67 | 2.085 | 1.025 |
| 1777-1804 | 36.10 | 28.36 | 26.12 | 22.48 |
| 1805-1831 | 10.51 | 6.155 | 3.145 | 0.9338 |
| 1832-1858 | 7.877 | 6.725 | 0.5424 | 2.024 |
| 1859-1885 | 0.5084 | 0.1805 | 8.370 | 4.331 |
| 1886-1912 | -0.5768 | -1.827 | 8.106 | 5.979 |
| 1913-1939 | 15.34 | 8.283 | 4.429 | 1.162 |
| 1940-1966 | 8.143 | 4.946 | -0.9441 | -2.984 |
| 1967-1993 | 2.848 | 1.605 | -1.580 | -0.4552 |
| 1994-2020 | 2.251 | -1.102 | 0.9984 | 2.241 |

| Period | VG | VG_DS | HVG_DT loc | VG_DT |
|-----------|-------|--------|------------|-------|
| 1753-1777 | 22.03 | 7.596 | 16.49 | 15.88 |
| 1777-1804 | 22.04 | 8.146 | 15.32 | 14.96 |
| 1805-1831 | 8.037 | 0.7347 | 6.760 | 6.568 |
| 1832-1858 | 4.543 | 3.429 | 4.438 | 6.134 |
| 1859-1885 | 7.679 | 1.472 | 6.335 | 3.484 |
| 1886-1912 | 4.070 | 3.996 | 3.486 | 3.905 |
| 1913-1939 | 13.82 | 2.890 | 9.412 | 9.313 |
| 1940-1966 | 10.04 | 4.178 | 9.346 | 7.871 |
| 1967-1993 | 4.888 | 1.711 | 2.215 | 3.616 |
| 1994-2020 | 8.961 | 0.5226 | 6.740 | 6.514 |

Table 4.4: Irreversibility ratio for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 10 time intervals. The coloring of the cells associates green to $IR \leq 1$ (reversibility), light green to $1 < IR \leq 4$ (reversibility with weak confidence), yellow to $4 < IR \leq 10$ (irreversibility with strong confidence) and orange to $IR > 10$ (irreversibility with extreme confidence).

Irreversibility ratio (IR)

| Period | HVG | HVG_DS | HVG_DT loc | HVG_DT |
|-----------|-------|---------|------------|--------|
| 1753-1804 | 114.3 | 75.12 | 15.63 | 15.44 |
| 1805-1858 | 35.80 | 27.40 | 1.949 | 1.417 |
| 1859-1912 | 3.156 | -0.9647 | 5.267 | 4.989 |
| 1913-1966 | 39.60 | 27.63 | 1.976 | 1.678 |
| 1967-2020 | 9.320 | 5.123 | 2.011 | 0.2669 |

| Period | VG | VG_DS | VG_DT loc | VG_DT |
|-----------|-------|--------|-----------|-------|
| 1753-1804 | 28.52 | 13.78 | 56.08 | 19.90 |
| 1805-1858 | 12.31 | 3.319 | 17.77 | 7.960 |
| 1859-1912 | 9.474 | 0.9784 | 16.97 | 7.777 |
| 1913-1966 | 17.56 | 3.866 | 37.38 | 11.45 |
| 1967-2020 | 9.094 | 2.029 | 16.54 | 5.943 |

Table 4.5: Irreversibility ratio for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 5 time intervals. The coloring of the cells associates green to $IR \leq 1$ (reversibility), light green to $1 < IR \leq 4$ (reversibility with weak confidence), yellow to $4 < IR \leq 10$ (irreversibility with strong confidence) and orange to $IR > 10$ (irreversibility with extreme confidence).

4.3.2 Motif detection

A motif detection algorithm is applied to the HVG computed over 5 and 10 intervals, the VG computed over 10 intervals and all their deseasoned counterparts; the most relevant frequency plots are presented in Fig. 4.13, and they correspond to motifs A, B and D. The motif frequencies associated to the deseasoned and original graphs are rather similar to each other, with point-wise discrepancies generally below 0.05; moreover for each frequency line the values tend to range in an interval of length smaller than 0.05. The partition of the series in 5 or 10 time intervals does not particularly affect the frequency plots for both the original and deseasoned graphs: for the HVGs motif A is the most frequent, followed closely by motif D, and both are always more frequent than motif B. On the other hand around 70% of the motifs detected for the VGs are of type D and for the original graph associated to the original series motif B is always more frequent than motif A.

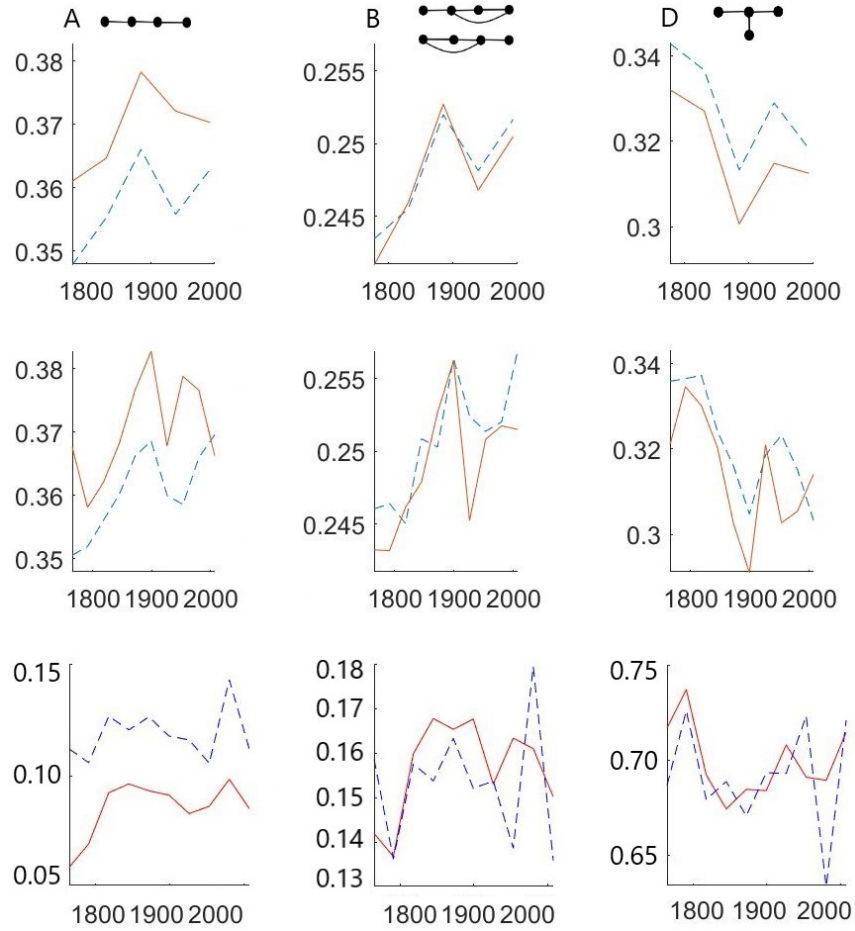


Figure 4.13: Frequencies of motifs A, B and D (left, center and right column respectively) for the HVGs computed over 5 intervals (top row), 10 intervals (center row) and for the VGs computed over 10 intervals (bottom row). The orange and red lines correspond to the original underlying time series, and the blue dashed line to its deseasoned counterpart.

Chapter 5

Case study 2: Prague

5.1 Series description

The series analysed in this section has been retrieved from the homogeneous blended European Climate Assessment dataset (ECA&D) and consists of daily maximum temperature measurements for the station of Praha-Klementinum in Czech Republic (50°05'11"N, 14°24'59"E). The series covers a period of 180 years, from the 1st of January 1825 to the 30th of April 2005.

5.1.1 Homogenization

The series in the ECA&D have been homogenized by an automated procedure described in detail in [23], that accounts for low availability or incompleteness of metadata, especially for the series further back in time. Nevertheless, metadata is necessary in instances like a simultaneous changes to the measurement networks at national scale, as both the target and reference series would be affected by the same break.

The breaks are detected at a yearly resolution with an agreement-based system based on three common methods (Prodige, RHtest and GAHMDI; for more information see section 2.1.2), meaning that a breakpoint is detected if at least two of these methods lead to the same result. The selection of reference series and combination of the detection methods are performed separately on annual and winter/summer half means of standardized differences between candidate and reference series; a maximum of 8 reference series is automatically selected on the basis of completeness, correlation of annual average (minimum 0.6) and distance (maximum 1000km). Breakpoints detected in adjacent years are considered as the same breakpoint, and at least three reference series must confirm a breakpoint in a pairwise

approach. The timing of the detected breaks (τ_1, \dots, τ_n) , from the most recent to the earliest, leads to the segmentation of the candidate into $n + 1$ sub-series $S_0(t|\tau_1 < t)$, $S_1(t|\tau_2 < t < \tau_1)$, etc. homogeneous by definition[3]. For each segment longer than 5 years, the adjustments are calculated by quantile matching on a monthly base and applied on a daily resolution.

The reference series are selected from a box of 6° centered on the candidate station and with an elevation difference smaller than 500 meters; in case of densely covered areas, the set union of the 40 longest ones and the 20 starting earliest is chosen. The break detection procedure is applied to the reference series to obtain homogeneous sub-series, and only those with at least 5 years of overlap with both segments of the candidate are selected; the maximum length of the sub-series is limited to 20 years. Finally, the 18 reference sub-series with highest daily raw correlation (> 0.75) with the segment of the candidate after the break are kept. In areas with a sparse network, up to 5 non-split series - meeting the correlation, geographical and temporal overlapping requirements - can be added to the reference set; in any case, a minimum of 3 reference series is required.

The breaks are considered in succession from τ_1 to τ_n . Given the break τ_i , the segment of the candidate after τ_i is termed the *basis* series, while the *segment* immediately before it is adjusted. For each month the distribution of temperatures is considered separately, introducing the seasonal cycle in the adjustments. A quantile sequence is generated for data before and after the break in the target $(s_{q,m}, b_{q,m})$ and in the reference series $(r_{j,q,m}^{bef}, r_{j,q,m}^{aft})$; given the target month m , also the absolute temperatures from the preceding and following month are considered in the calculation of the quantiles to reduce the noise. The adjustments are calculated as

$$a_{i,j,q,m} = (b_{q,m} - s_{i,q,m}) - (r_{j,q,m}^{aft} - r_{j,q,m}^{bef}), \quad (5.1)$$

and they are smoothed by considering the mean of adjustments from neighboring months and quantiles:

$$\bar{a}_{j,q,m} = (a_{j,q,m} + a_{j,q+5,m} + a_{j,q-5,m} + a_{j,q,m+1} + a_{j,q,m-1})/5. \quad (5.2)$$

A set of estimations of the correction is produced, each one corresponding to the different overlapping periods each reference series R_j has with the segments of the candidate. The value to be corrected may belongs to a different quantile \tilde{q}_j in each of the overlapping periods, so the estimation of the adjusted value related to R_j is $\tilde{v}_j = v + a_{j,\tilde{q}_j,m}$, where v is the original value. The final adjusted value is the median of the estimations \tilde{v}_j .

5.1.2 Missing values

The non-blended series is complete, but as a result of the homogenization process the corresponding blended series features 22 missing values. These are distributed in an interval that ranges from 1827 to 1956; 13 years lack only one entry, the years 1922 and 1925 lack two entries each and the year 1920 lacks 5. Moreover the first three missing values in 1920 are concentrated in January, which results in the shortest time differences between missing dates in the series: 17 days with the missing value in 1919, 11 days between the first two missing values of 1920 and 6 between the second and third. Every other pair of missing dates has at least 30 days of distance from each other. The series is completed by assigning to each missing value the average between the previous and following temperature record.

5.1.3 Preliminary analysis

The maximum temperature ranges from -20.5°C to 37.8°C , recorded on January 22 1850 and July 27 1983 respectively. The series can be fitted with a bimodal distribution, as shown in Fig.5.1; the peaks for the maximum temperature distribution are 5 and 20°C .

The earliest maximum temperature is recorded on April 16, 2005, followed by another instance in May 1847; every other yearly maximum temperature is recorded between June and August, with occurrences most frequent in July.

The monthly average and standard deviation of maximum temperatures, divided between values before and after 1901, are presented in Fig.5.2. The monthly average plot is qualitatively similar to Fig.4.4: January is the coldest month, and July the hottest. The values are lower for the most recent period, with the exception of September and October for the maximum temperature. The standard deviation instead features lower variability between the months for both intervals, and lower difference between each pair of monthly values; moreover most of the highest values of standard deviation are associated to the period 1901-2004.

The yearly average of maximum temperature ranges from 11.2°C , recorded in 1838, to 15.9°C , recorded in 1834. Their values and the 9-year moving average are reported in Fig.5.3. The data follows a positive linear trend, with increasing slope for the more recent intervals.

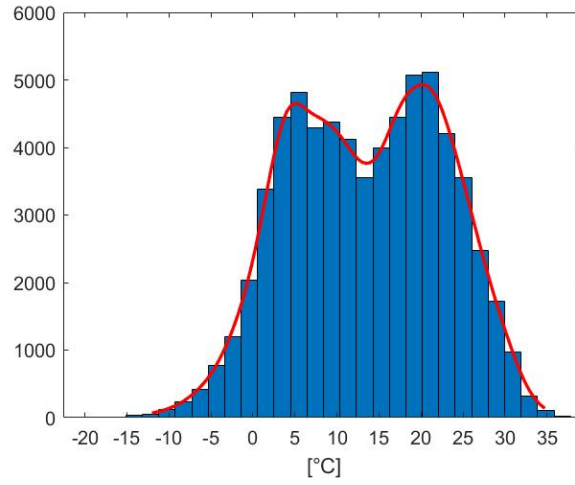


Figure 5.1: Relative frequency (%) of daily maximum temperatures. The red curve is the bimodal distribution fitted to the histogram.

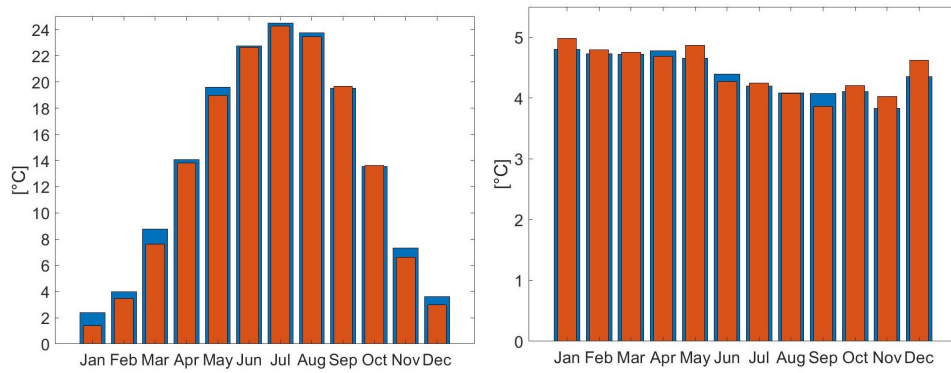


Figure 5.2: Monthly mean (left) of maximum temperatures and corresponding standard deviation (right). The blue series refers to the period 1825-1900, and the orange one to 1901-2004.

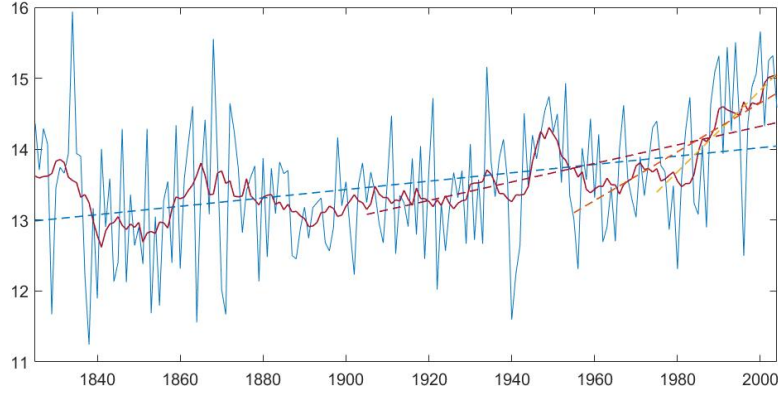


Figure 5.3: Yearly average of maximum temperature (blue line) and 9-year moving average (red line); the linear trends computed over the whole period, the last 100, 50 and 30 years of the series are highlighted in a dashed blue, red, orange and yellow line respectively.

One of the consequences of the increasing maximum temperature is the reduction of number of frost days in a year, as shown in Fig.5.4. The year with the highest number of days with maximum temperature below 0°C is 1838, and the record of frost days gradually lowers to 47 in 1996. The plot highlights a reduction in the number of frost days of almost 37% over the 180 year period.

On the other hand, the increase in number of days with maximum temperature above 30°C is less obvious; as shown in Fig.5.5, the value of the peaks slightly decreases over time, from 30 days in 1834 to 27 in 1994 and 2003. The local minima instead show a more clear increase in value in the last decades of the series: the last year with no hot days is 1956 and the minimum of hot days in the following years increases to 4 in 1996 and to 10 in 2004.

Differently from the previous case study, though, the number of heat waves detected is significantly lower, as shown in Fig.5.6; this implies that the majority of years in the series do not feature periods of at least 6 days with maximum temperature $\geq 28^{\circ}\text{C}$, and that even when hot days occur they are generally followed by colder days. Nevertheless, both the record of heat wave intensity (34.95°C) and length (20 days) occur in the last decade of the series, in 1994 and 2003 respectively.

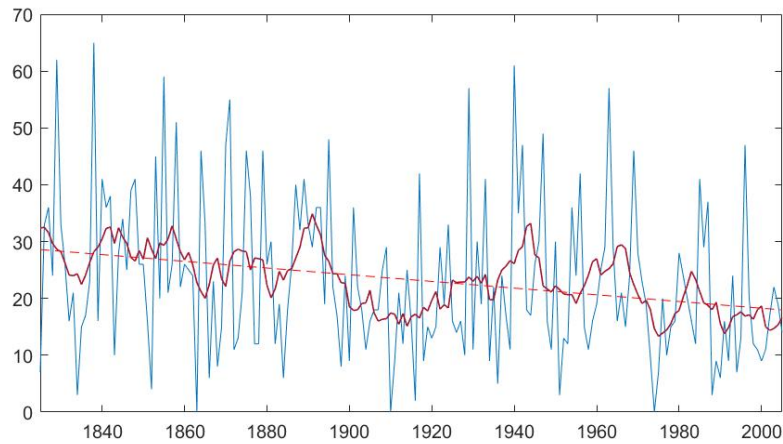


Figure 5.4: Number of frost days in a year (blue line), smoothed average with a 9-year window (red line) and linear trend (dashed bright red line).

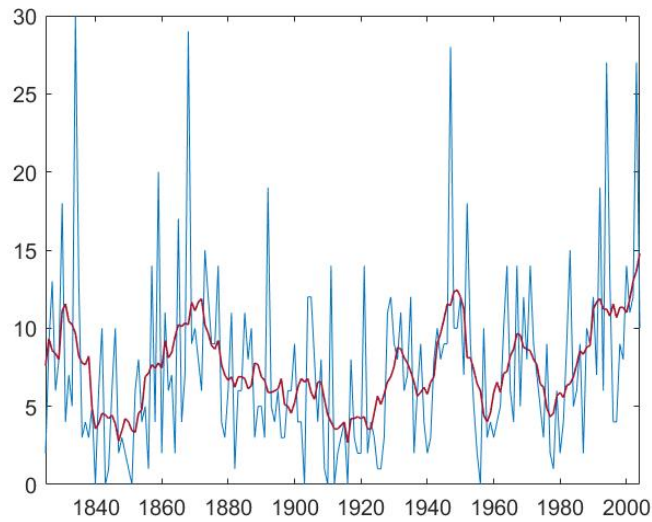


Figure 5.5: Number of days in a year with maximum temperature above 30°C (blue line), and smoothed average with a 9-year window (red line).

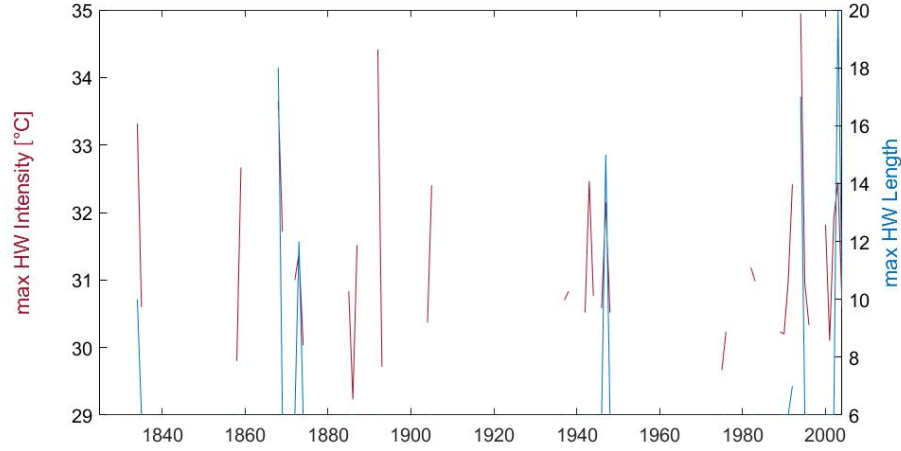


Figure 5.6: Maximum intensity of the heat waves detected each year, as defined by Mercalli [17], and maximum length of the heat wave, i.e. number of consecutive days with maximum daily temperatures $\geq 28^\circ$.

5.2 Visibility graphs

5.2.1 Degree metrics

The first four moments of the degree for the HVG, VG and their detrended and deseasoned counterparts computed on the full series are presented in table 5.1. Interestingly, all the observations conducted in section 4.3.1 in relation to table 4.3 apply in this case as well. All the metrics for the VGs are greater than the corresponding value for the HVGs; in particular the values of standard deviation and kurtosis for the VGs are more than double the corresponding value the HVGs. Moreover the difference in standard deviation and kurtosis between the deseasoned and original VG is greater in absolute value than the corresponding difference for the HVGs.

The degree moments for the graphs computed on 4 and 7 intervals are shown in figure 5.7. Each metric is characterized by a significant linear trend, but for some instances the slope is positive if the series is partitioned in 4 intervals, and negative otherwise. This effect is most likely induced by the low number of intervals and should not have any real implication on the interpretation of the result: the data points of the red and blue series for each plot are indeed close in value. This occurs in the degree skewness of all versions of HVG, but it can also be observed for the standard

Degree Metrics

| Graph | Mean | Std dev | Skewness | Kurtosis |
|----------|-------|---------|----------|----------|
| HVG | 3.902 | 1.921 | 1.530 | 6.364 |
| HVG (DT) | 3.999 | 2.006 | 1.542 | 6.439 |
| HVG (DS) | 3.999 | 2.030 | 1.642 | 7.180 |
| VG | 9.314 | 9.235 | 2.788 | 13.89 |
| VG (DT) | 9.324 | 9.240 | 2.786 | 13.88 |
| VG (DS) | 7.877 | 6.185 | 2.950 | 20.19 |

Table 5.1: First four moments of the degree for the HVG, VG and their detrended (DT) and deseasoned (DS) counterparts, computed over the full series.

deviation, skewness and kurtosis of the deseasoned VG, the mean and kurtosis of the deseasoned HVG, and the standard deviation and kurtosis of the HVG. The sign of the slope for the remaining plots corresponds to figure 4.10, with the exception of the mean degree of the deseasoned HVG and the kurtosis of the detrended HVG, which feature a positive linear trend.

5.2.2 Assortativity

The neighborhood connectivity of the HVGs and VGs computed over 4 and 7 intervals are reported in figures 5.8 and 5.9 respectively. The observations conducted in section 4.3.1 still apply: in particular the plot points lie above the bisector only for degree values lower than 4 in the case of the HVGs, and 13 for the deseasoned VG. For the VG and its detrended counterparts, instead, the bisector is crossed for degree values between 23 and 27 depending on the time interval considered; the most recent series intersect the bisector for lower degree values, and the least recent intersect for higher degree values. The major difference with the previous case study is that for all plots associated to the VGs the points of different intervals are less scattered.

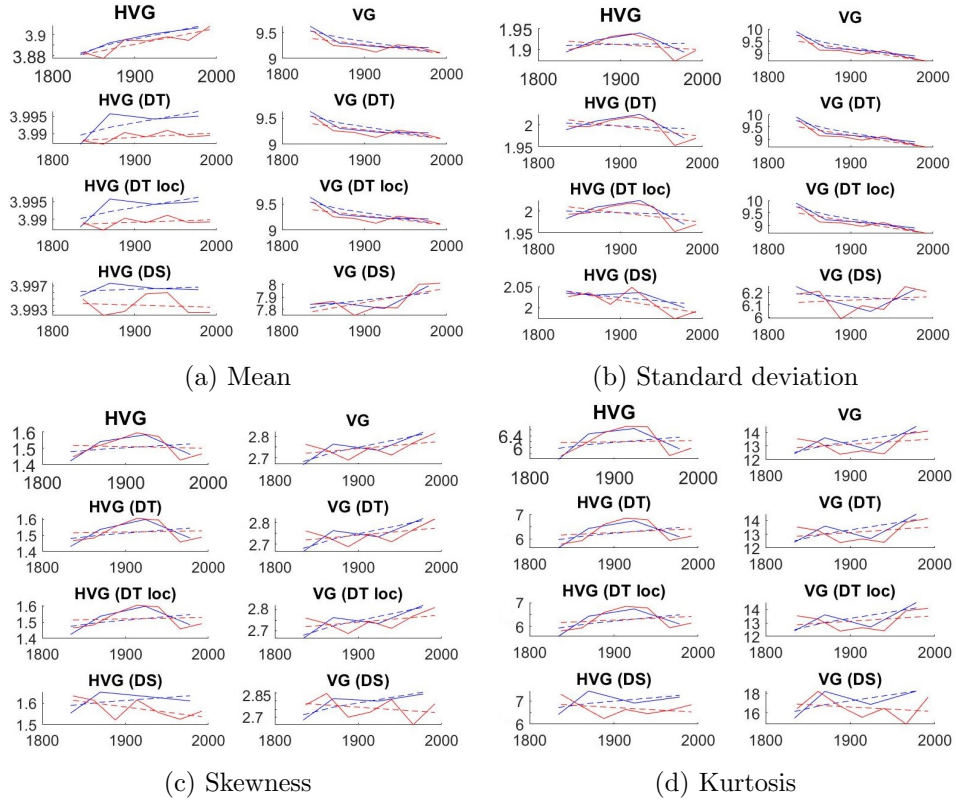


Figure 5.7: First four moments of the degree for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 5 intervals (blue line) and 10 intervals (red line). The corresponding linear trends are highlighted with a dashed line when significant.

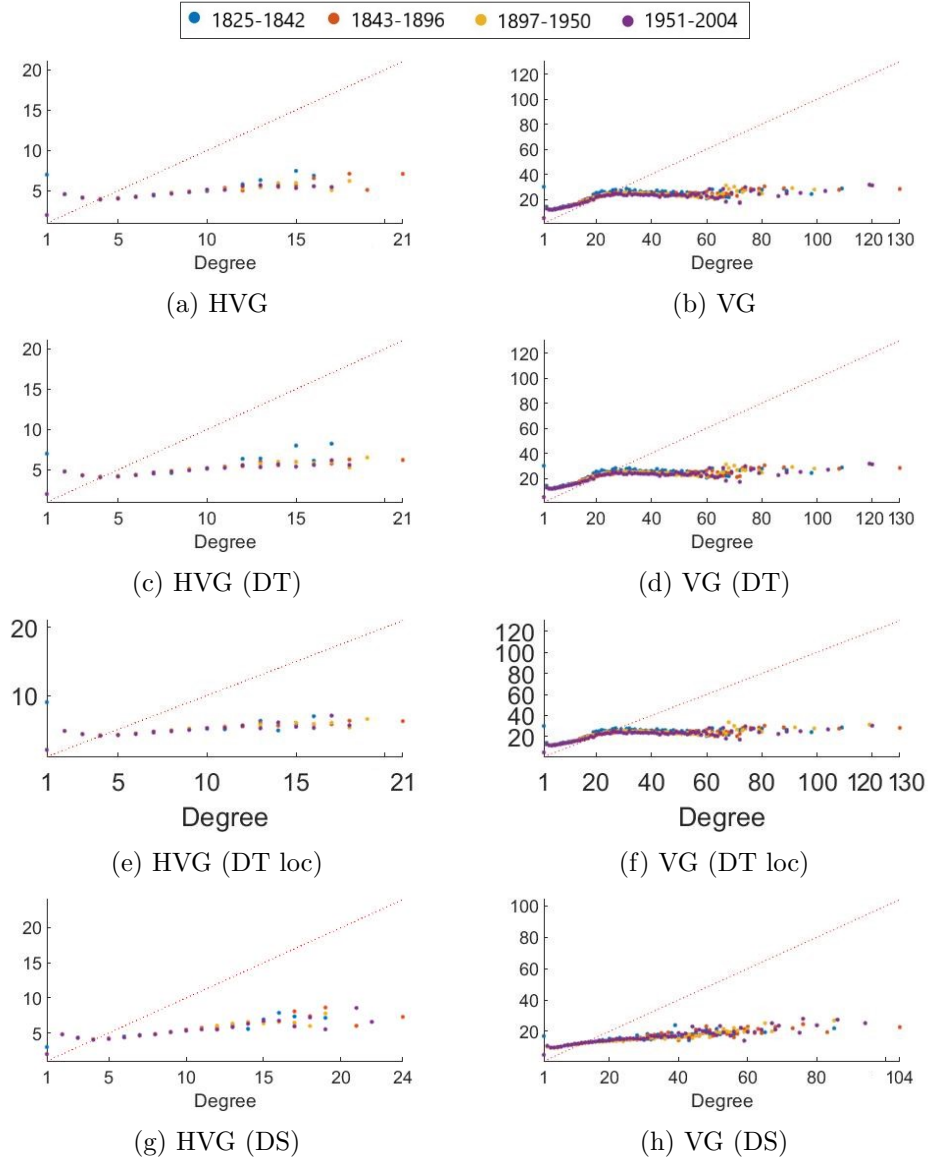


Figure 5.8: Neighborhood connectivity for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 4 intervals. The bisector $y = x$ is highlighted with a red dashed line for each plot.

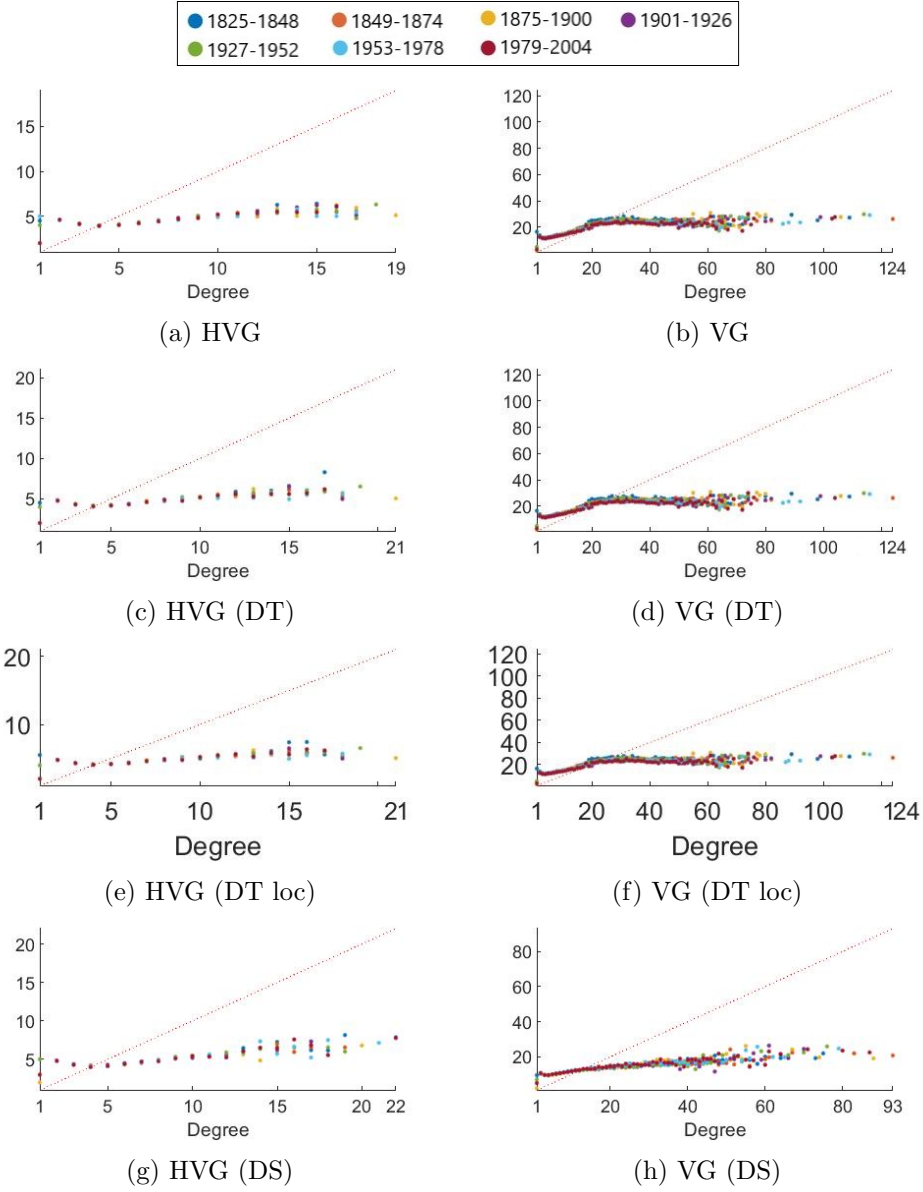


Figure 5.9: Neighborhood connectivity for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 7 intervals. The bisector $y = x$ is highlighted with a red dashed line for each plot.

5.2.3 Time reversibility

The irreversibility ratios relative to the HVGs and VGs computed over 7 and 4 intervals are presented in table 5.3. Each HVG, VG and detrended VG lead to a value of IR greater than 4 on each interval of both partitions, meaning that they are all irreversible; also the deseasoned HVG results in mostly irreversible intervals. The deseasoned VG results in irreversible intervals on the 4-interval partition, but on the finer one 3 intervals are labeled reversible with weak confidence. The most surprising result though is given by the detrended HVG: the 4-interval partition highlights only 1951-2004 as a reversible period, but in the finer partition the 3 intervals covering the period 1927-2004 are reversible and every other interval is reversible with at least weak confidence.

| Irreversibility ratio (IR) | | | | |
|----------------------------|-------|--------|------------|----------|
| Period | HVG | HVG_DS | HVG_DT loc | HVG_DT |
| 1825-1848 | 16.75 | 6.097 | 45.74 | 3.003 |
| 1849-1874 | 26.12 | 11.00 | 7.599 | 2.086 |
| 1875-1900 | 15.87 | 6.632 | 4.968 | 0.4917 |
| 1901-1926 | 16.31 | 6.799 | 3.983 | 2.556 |
| 1927-1952 | 7.074 | 3.892 | 1.846 | -1.368 |
| 1953-1978 | 10.07 | 5.349 | 0.3765 | -0.04739 |
| 1979-2004 | 5.405 | 5.272 | 0.2034 | 0.3669 |

| Period | VG | VG_DS | VG_DT loc | VG_DT |
|-----------|-------|-------|-----------|-------|
| 1825-1848 | 9.888 | 5.340 | 15.77 | 11.39 |
| 1849-1874 | 11.73 | 8.902 | 16.57 | 17.11 |
| 1875-1900 | 9.785 | 2.769 | 15.09 | 11.76 |
| 1901-1926 | 6.877 | 5.542 | 10.71 | 11.05 |
| 1927-1952 | 6.113 | 5.179 | 10.16 | 8.011 |
| 1953-1978 | 4.647 | 3.927 | 10.43 | 8.962 |
| 1979-2004 | 7.581 | 3.047 | 10.44 | 9.955 |

Table 5.2: Irreversibility ratio for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 7 intervals. The coloring of the cells associates green to $IR \leq 1$ (reversibility), light green to $1 < IR \leq 4$ (reversibility with weak confidence), yellow to $4 < IR \leq 10$ (irreversibility with strong confidence) and orange to $IR > 10$ (irreversibility with extreme confidence).

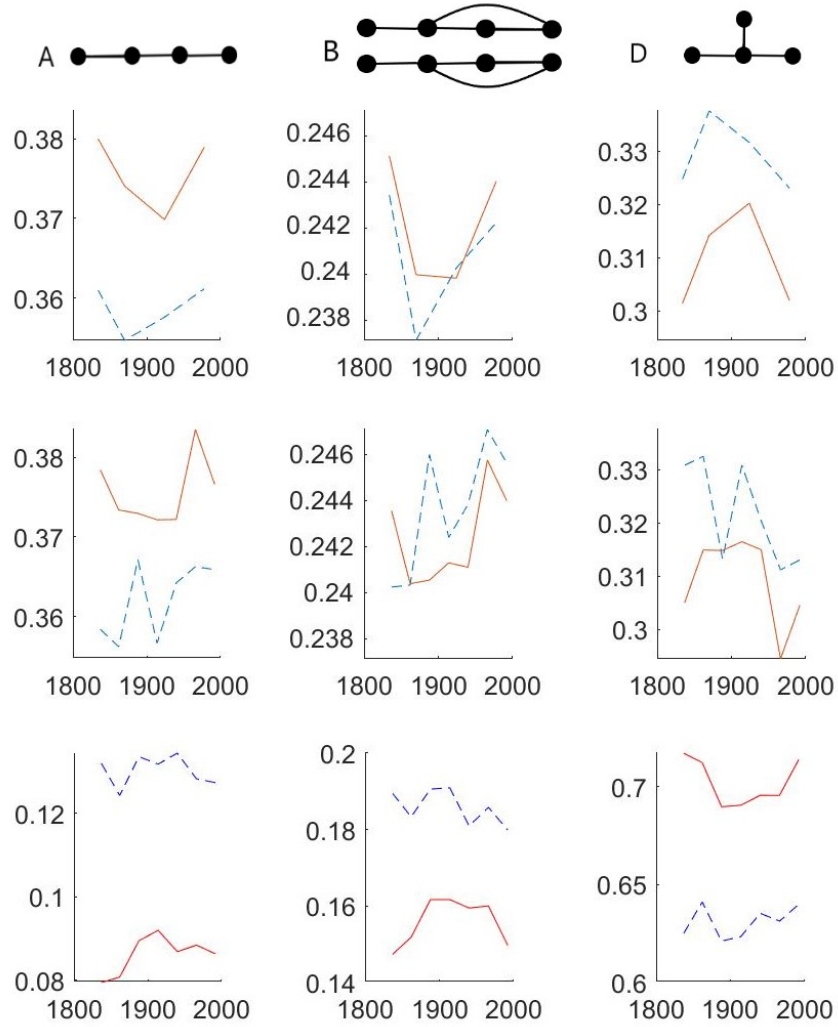
| Irreversibility ratio (IR) | | | | |
|----------------------------|-------|--------|------------|--------|
| Period | HVG | HVG_DS | HVG_DT loc | HVG_DT |
| 1825-1842 | 8.569 | 6.069 | 54.15 | 3.165 |
| 1843-1896 | 26.89 | 31.76 | 15.43 | 8.891 |
| 1897-19050 | 23.20 | 21.57 | 17.51 | 6.831 |
| 1951-2004 | 10.25 | 9.336 | 0.5682 | 0.6910 |

| Period | HVG | HVG_DS | HVG_DT loc | HVG_DT |
|------------|-------|--------|------------|--------|
| 1825-1842 | 15.68 | 4.477 | 6.872 | 16.11 |
| 1843-1896 | 19.82 | 12.91 | 12.04 | 18.23 |
| 1897-19050 | 19.52 | 14.50 | 9.827 | 17.25 |
| 1951-2004 | 15.11 | 13.42 | 10.51 | 15.05 |

Table 5.3: Irreversibility ratio for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 4 intervals. The coloring of the cells associates green to $IR \leq 1$ (reversibility), light green to $1 < IR \leq 4$ (reversibility with weak confidence), yellow to $4 < IR \leq 10$ (irreversibility with strong confidence) and orange to $IR > 10$ (irreversibility with extreme confidence).

5.2.4 Motif detection

A motif detection algorithm is applied to the HVG computed over 4 and 7 intervals, the VG computed over 7 intervals and all their deseasoned counterparts; the most relevant frequency plots are presented in Fig. 5.10. Interestingly, the results obtained have many features in common with the frequency plots discussed in section 4.3.2, especially for the HVGs. The greatest difference is observed for the frequency plots of motifs B and D of the deseasoned VG, as the corresponding plot lines do not intersect the lines associated to the original graph.



Chapter 6

Case study 3: Bologna

6.1 Series description

The data analyzed in this section is another homogeneous blended¹ series from the European Climate Assessment dataset (ECA&D), that refers to daily maximum temperature measurements for the station of Bologna, in Italy. The series spans over 190 years, from the 1st of January 1814 to the 31st of December 2003, and is characterized by 140 missing values as a result of the homogenization process. There is a total of 64 years featuring missing data points, and about 66% of them are concentrated in the periods 1814-1841 and 1950-1976; there are at most 5 missing entries in the same year and there are only 4 instances on the whole series with two consecutive missing dates. This allows to perform data imputation by assigning to each missing value the average of the previous and following temperature in the series.

6.1.1 Preliminary analysis

The maximum temperature ranges from -8.2°C to 40.7°C , recorded on February 13 1929 and July 11 1870 respectively. The series can be fitted with a bimodal distribution, as shown in Fig.6.1; the peaks for the maximum temperature distribution are 8 and 28°C , with a minor peak at 19°C .

The earliest yearly maximum temperature is recorded on May 23, 1847; the other maxima are distributed between June and August, with about 54% of occurrences in July, 33% in August and 12% in June.

¹For information on the homogenization process refer to section 5.1.1.

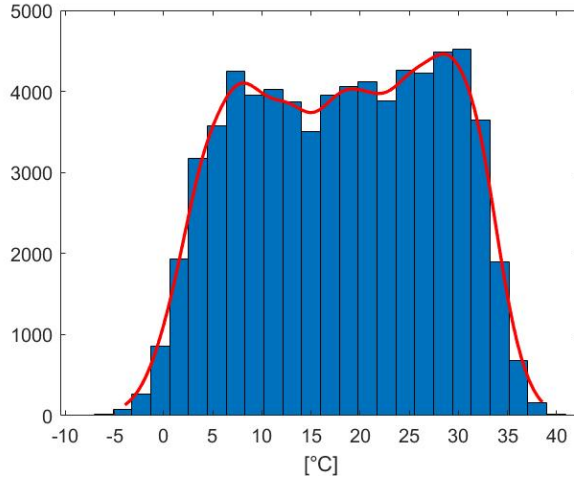


Figure 6.1: Relative frequency (%) of daily maximum temperatures. The red curve is the bimodal distribution fitted to the histogram.

The monthly average and standard deviation of maximum temperatures, divided between values before and after 1901, are presented in Fig.6.2. Compared to the interval 1814-1900, the period 1901-2003 features higher averages from April to October and lower averages in the other months. The standard deviation plot is similar both qualitatively and quantitatively to the one reported in Fig.5.2, with the lowest values recorded in September and November.

The yearly average of maximum temperature ranges from 16.0°C, recorded in 1850, to 21.2°C, recorded in 2000. Their values and the 9-year moving average are reported in Fig.6.3; the smoothed data shows periodic oscillations, with the highest peak reached between 1867 and 1878. The data follows a positive linear trend, with increasing slope for the more recent intervals.

The increase in temperature reduces significantly the number of days in a year with daily maximum temperature below 0°C, as shown in Fig.6.4. The highest value is 26 days in 1830, and the following peaks decrease gradually in value for over a century; over the last 50 years of the series the peaks drop from 18 days in 1963 to 10 in 1985, and no other local minimum above 5 days is recorded afterwards.

The number of hot days in a year is presented in Fig.6.5; the smoothed

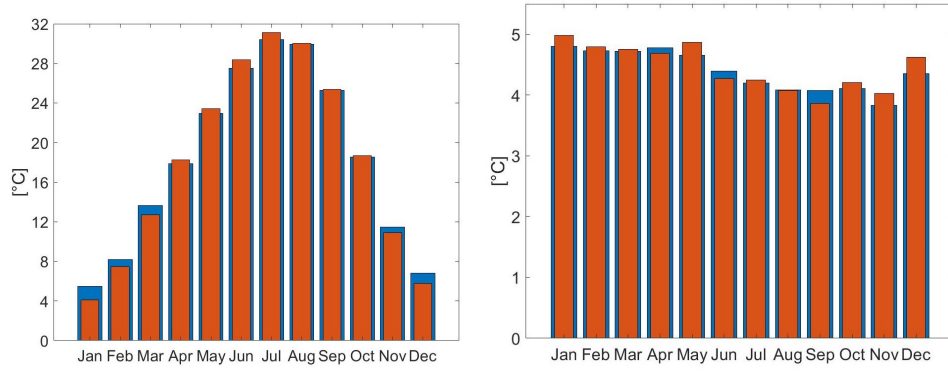


Figure 6.2: Monthly mean (left) of maximum temperatures and corresponding standard deviation (right). The blue series refers to the period 1814-1900, and the orange one to 1901-2003.

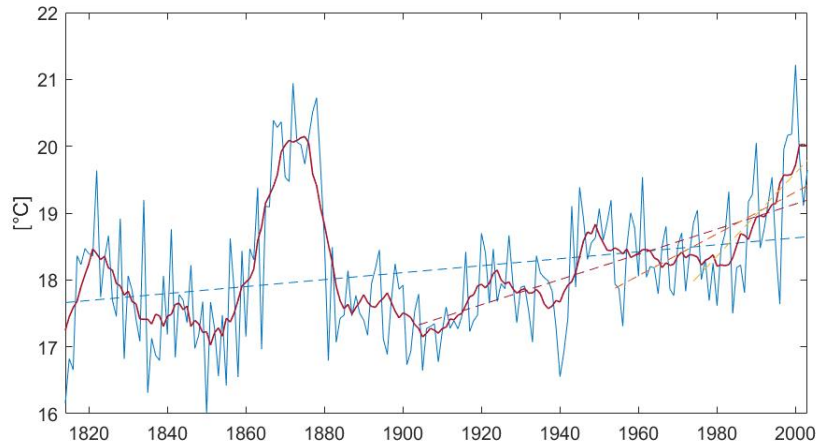


Figure 6.3: Yearly average of maximum temperature (blue line) and 9-year moving average (red line); the linear trends computed over the whole period, the last 100, 50 and 30 years of the series are highlighted in a dashed blue, red, orange and yellow line respectively.

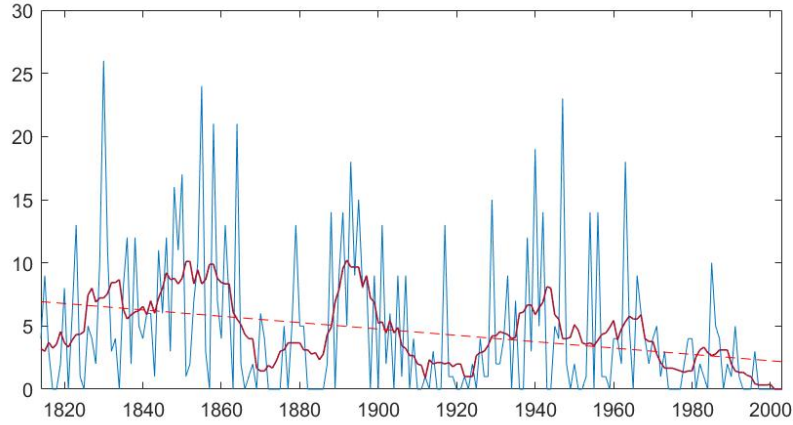


Figure 6.4: Number of frost days in a year (blue line), smoothed average with a 9-year window (red line) and linear trend (dashed red line).

average in particular highlights the peak between 1860 and 1879 and the gradual increase of the last century of the series. The maximum number of days in a year with maximum temperature above 30°C is 100 in 1877; the last 50 years of the series feature a maximum of 81 hot days in 1998, and a minimum that increases from 23 in 1968 to 33 in 2002.

Differently from what observed in the previous case studies, the series always feature heat waves, as shown in Fig.6.6. The lowest intensity of heat wave is 29.43°C in 1940, and it also corresponds to the lowest duration of 6 days; the maximum duration and intensity of heat wave, instead, are recorded in 1873 and 2000 respectively. The local minima and maxima increase slightly in the last 60 years of the series, but it is harder to define a linear trend for either the intensity or intensity of the heat waves.

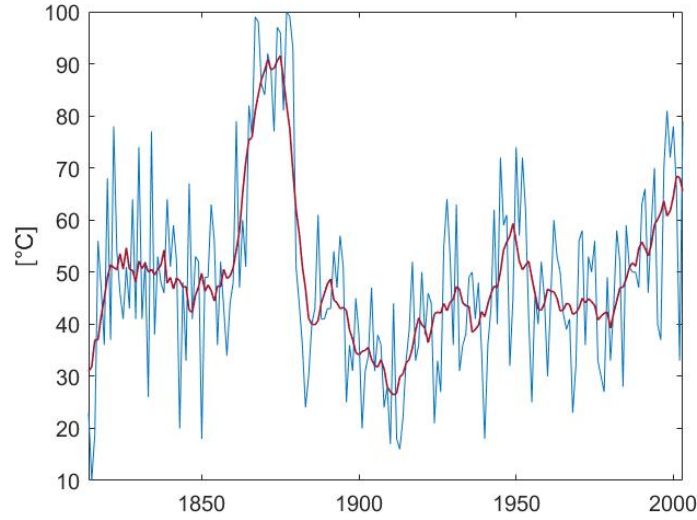


Figure 6.5: Number of days in a year with maximum temperature above 30°C (blue line), and smoothed average with a 9-year window (red line).

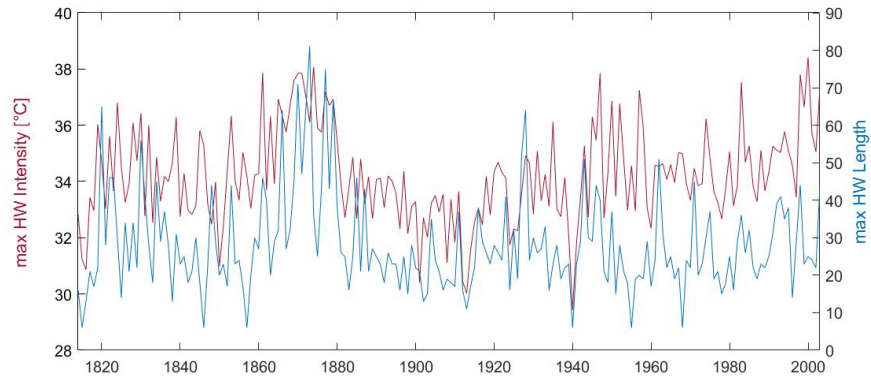


Figure 6.6: Maximum intensity of the heat waves detected each year, as defined by Mercalli [17], and maximum length of the heat wave, i.e. number of consecutive days with maximum daily temperatures $\geq 28^{\circ}$.

Degree Metrics

| Graph | Mean | Std dev | Skewness | Kurtosis |
|----------|-------|---------|----------|----------|
| HVG | 3.842 | 1.921 | 1.628 | 6.982 |
| HVG (DT) | 3.998 | 2.055 | 1.612 | 6.824 |
| HVG (DS) | 3.999 | 2.085 | 1.723 | 7.554 |
| VG | 10.08 | 11.13 | 2.602 | 11.41 |
| VG (DT) | 10.06 | 11.06 | 2.597 | 11.44 |
| VG (DS) | 7.707 | 6.563 | 3.659 | 31.00 |

Table 6.1: First four moments of the degree for the HVG, VG and their detrended (DT) and deseasoned (DS) counterparts, computed over the full series.

6.2 Visibility graphs

6.2.1 Degree metrics

The first four moments of the degree for the HVG, VG and their detrended and deseasoned counterparts computed on the full series are presented in table 6.1. All the metrics for the VGs are greater than the corresponding value for the HVGs; the relative difference between a graph and its deseasoned counterpart is greater for the VG than the HVG. For example the degree kurtosis of the deseasoned VG is more than double the corresponding value for the VG and its detrended counterpart. The mean and standard deviation of the HVG are slightly lower than the corresponding values for the detrended HVG, and these are lower w.r.t. the deseasoned HVG; the same applies to the kurtosis of the VGs. On the other hand, the mean and standard deviation of the deseasoned VG are lower than the values for the detrended VG, and these are lower w.r.t. the VG.

The degree moments for the graphs computed on 4 and 7 intervals are shown in figure 6.7. Each metric is characterized by a significant linear trend, that is positive for the mean and standard deviation of the HVGs, the skewness and kurtosis of the VGs and negative otherwise. There are 3 instances where the slope differs in sign depending on the partition considered, specifically the mean of the deseasoned HVG and the skewness and kurtosis of the deseasoned VG; nevertheless, the data points of the red and blue series for each plot are very close in value.

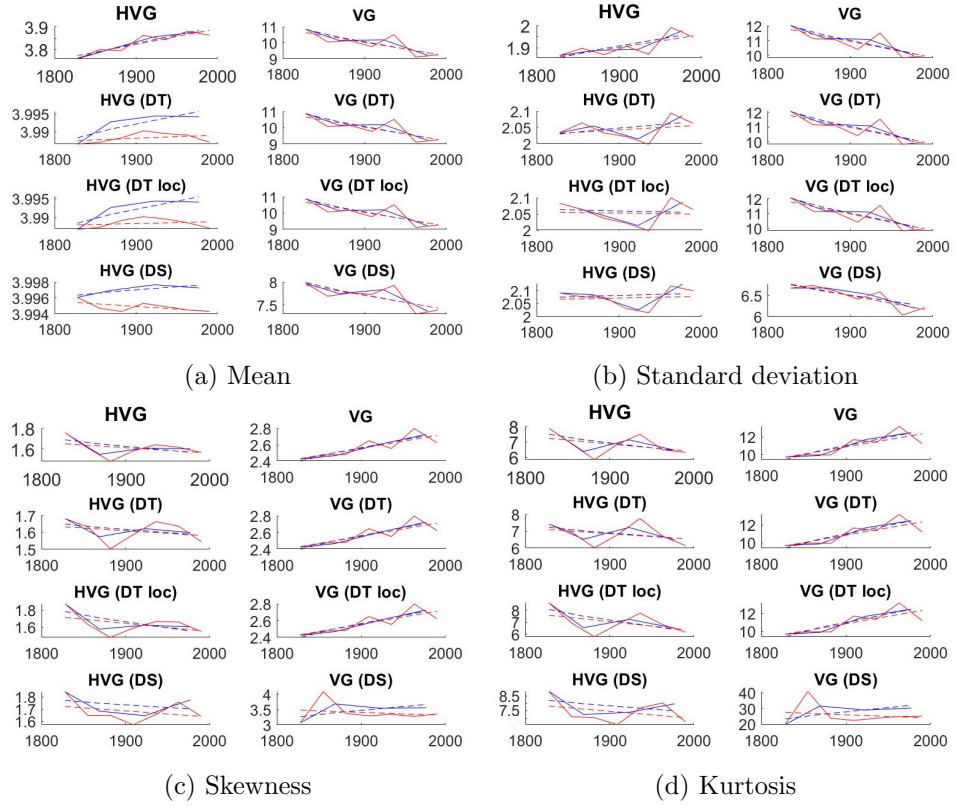


Figure 6.7: First four moments of the degree for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 5 intervals (blue line) and 10 intervals (red line). The corresponding linear trends are highlighted with a dashed line when significant.

6.2.2 Assortativity

The neighborhood connectivity of the HVGs and VGs computed over 4 and 7 intervals are reported in figures 6.8 and 6.9 respectively. The plots correspond qualitatively to those discussed in section 5.2.2; in particular those related to the HVGs are also very similar quantitatively. The largest difference between the partitions is found in the neighborhood connectivity plot of the VG: the data points of the oldest interval cross the bisector for lower degree values for the graphs computed over 10 intervals.

6.2.3 Time reversibility

The irreversibility ratios relative to the HVGs and VGs computed over 7 and 4 intervals are presented in table 6.3. The VG and its detrended counterpart lead to irreversibility on every interval of both partitions considered; the same applies to the deseasoned HVG as well. Similarly the HVG leads to intervals that are irreversible with extreme confidence, with the exception of 1923-1949 that is reversible with weak confidence. The series is mostly irreversible also w.r.t. the deseasoned VG; only one interval for each partition is reversible with weak confidence, but there is no overlap between them. Reversible intervals are detected only in relation to the detrended HVG, and they cover the period 1842-1949 in both partitions.

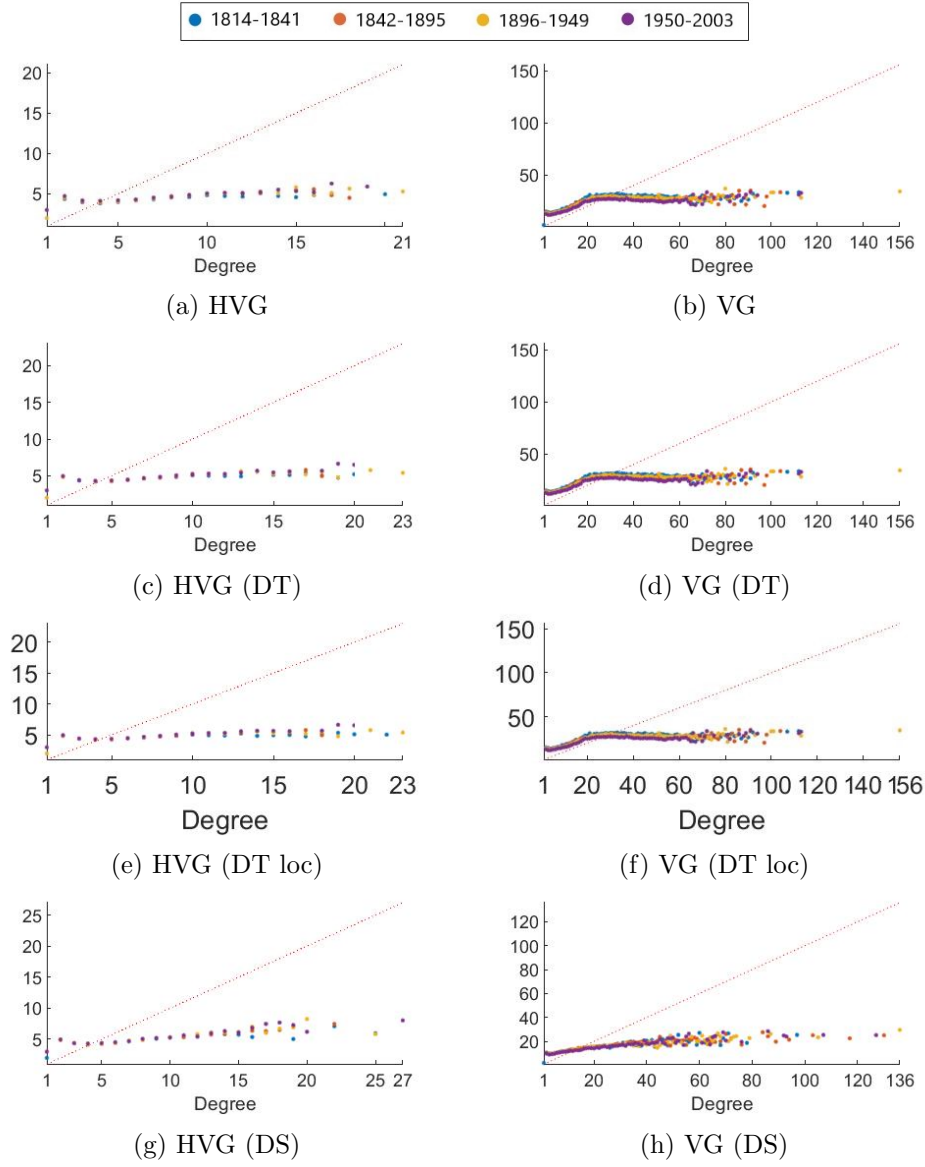


Figure 6.8: Neighborhood connectivity for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 5 intervals. The bisector $y = x$ is highlighted with a red dashed line for each plot.

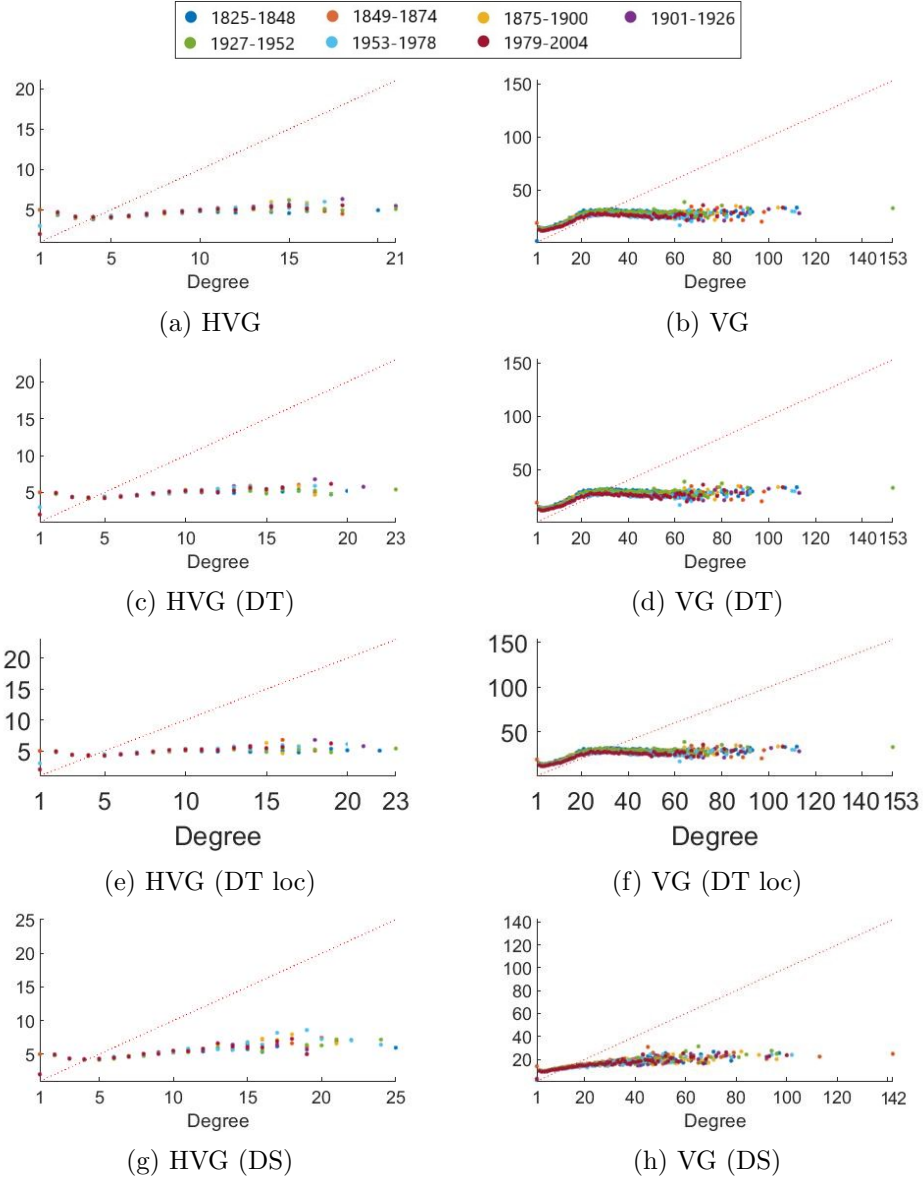


Figure 6.9: Neighborhood connectivity for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 10 intervals. The bisector $y = x$ is highlighted with a red dashed line for each plot.

Irreversibility ratio (IR)

| Period | HVG | HVG_DS | HVG_DT loc | HVG_DT |
|-----------|-------|--------|------------|----------|
| 1814-1841 | 82.22 | 50.90 | 147.9 | 16.62 |
| 1842-1868 | 12.89 | 5.205 | 0.5631 | 0.8165 |
| 1869-1895 | 15.95 | 10.51 | 48.34 | -3.186 |
| 1896-1922 | 13.92 | 8.386 | -0.9912 | -0.08822 |
| 1923-1949 | 3.835 | 4.612 | -1.981 | 0.3315 |
| 1950-1976 | 39.86 | 17.17 | 44.24 | 24.74 |
| 1977-2003 | 36.46 | 22.78 | 12.33 | 24.35 |

| Period | VG | VG_DS | VG_DT loc | VG_DT |
|-----------|-------|-------|-----------|-------|
| 1814-1841 | 41.10 | 30.43 | 24.22 | 47.91 |
| 1842-1868 | 12.80 | 5.537 | 5.250 | 14.55 |
| 1869-1895 | 11.65 | 7.653 | 8.889 | 17.09 |
| 1896-1922 | 13.56 | 7.031 | 8.028 | 16.18 |
| 1923-1949 | 8.234 | 2.632 | 5.953 | 9.528 |
| 1950-1976 | 19.91 | 10.38 | 12.64 | 25.02 |
| 1977-2003 | 20.59 | 8.997 | 14.25 | 24.92 |

Table 6.2: Irreversibility ratio for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 7 intervals. The coloring of the cells associates green to $IR \leq 1$ (reversibility), lime to $1 < IR \leq 4$ (reversibility with weak confidence), yellow to $4 < IR \leq 10$ (irreversibility with strong confidence) and orange to $IR > 10$ (irreversibility with extreme confidence).

| Irreversibility ratio (IR) | | | | |
|----------------------------|-------|--------|------------|--------|
| Period | HVG | HVG_DS | HVG_DT loc | HVG_DT |
| 1814-1841 | 116.1 | 61.51 | 341.9 | 8.319 |
| 1842-1895 | 34.91 | 18.80 | -2.170 | -1.110 |
| 1896-1949 | 22.61 | 20.29 | -5.148 | -2.644 |
| 1950-2003 | 103.5 | 56.85 | 53.38 | 24.98 |

| Period | VG | VG_DS | VG_DT loc | VG_DT |
|-----------|-------|-------|-----------|-------|
| 1814-1841 | 15.96 | 17.98 | 27.48 | 25.42 |
| 1842-1895 | 6.238 | 3.937 | 14.48 | 13.94 |
| 1896-1949 | 6.196 | 5.579 | 8.738 | 10.99 |
| 1950-2003 | 12.80 | 12.62 | 22.70 | 22.93 |

Table 6.3: Irreversibility ratio for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 4 intervals. The coloring of the cells associates green to $IR \leq 1$ (reversibility), lime to $1 < IR \leq 4$ (reversibility with weak confidence), yellow to $4 < IR \leq 10$ (irreversibility with strong confidence) and orange to $IR > 10$ (irreversibility with extreme confidence).

6.2.4 Motif detection

A motif detection algorithm is applied to the HVG computed over 4 and 7 intervals, the VG computed over 7 intervals and all their deseasoned counterparts; the most relevant frequency plots are presented in Fig. 6.10. As already observed in the previous examples, motifs D and A are more frequent for the HVGs, whereas motif D is the most frequent in the VG; moreover the range of each motif frequency is similar to that of the other case studies.

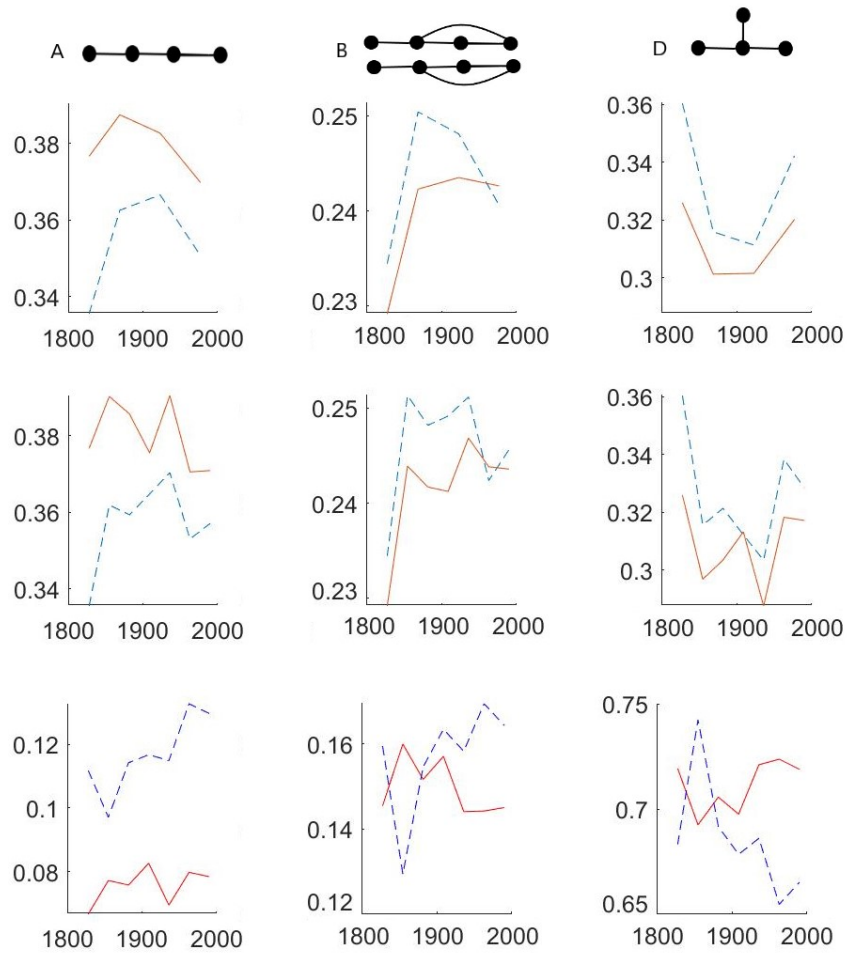


Figure 6.10: Frequencies of motifs A, B and D (left, center and right column respectively) for the HVGs computed over 4 intervals (top row), 7 intervals (center row) and for the VGs computed over 7 intervals (bottom row). The orange and red lines correspond to the original underlying time series, and the blue dashed line to its deseasoned counterpart.

Chapter 7

Discussion and Conclusions

7.1 Case study comparisons

The main results of the case studies are here summarized and compared. As shown in Tab. 7.1, the daily maximum temperatures recorded in Prague are generally lower than those of Turin and Bologna: the coldest and hottest temperatures are lower, and so are the peaks of relative frequency. For each series the yearly average of maximum temperatures is characterized by a positive linear trend that is particularly evident in the most recent decades of observations; this coincides with a significant reduction in the number of frost days per year. On the other hand the heat wave analysis shows only for the Turin series a clear increase in heat wave intensity and duration in recent years.

For each visibility graph computed, the first three moments of the degree have similar values, as shown in tables 7.1 and 7.2. Moreover the metrics associated to a HVG and its detrended and deseasoned counterparts are very close in value, and the same applies to a VG and its detrended counterpart.

The neighborhood connectivity plots display very similar qualitative and quantitative features for all the case studies: the majority of points lie below the bisector for values of degree greater or equal to 5 for the HVGs and 13 for the deseasoned VGs; for the VGs and their detrended counterparts the bisector is crossed for values of degree between 21 and 30 depending on the interval of the series considered. The highest level of disassortativity for high degree nodes is observed for the deseasoned VGs, and the points belonging to different intervals are a lot more scattered

for the VGs and detrended VGs on the Turin series. It is important to note that the series points that are most scattered belong to the oldest intervals, which coincidentally were the most affected by missing values and inhomogeneities. An example of the neighborhood connectivity plots is reported in Fig. 7.1 for reference.

| | Mean | | | | Standard deviation | | |
|----------|-------|-------|-------|--|--------------------|-------|-------|
| Graph | T | P | B | | T | P | B |
| HVG | 3.898 | 3.902 | 3.842 | | 2.035 | 1.921 | 1.921 |
| HVG (DT) | 3.999 | 3.999 | 3.998 | | 2.133 | 2.006 | 2.005 |
| HVG (DS) | 3.999 | 3.999 | 3.999 | | 2.147 | 2.030 | 2.085 |
| VG | 9.163 | 9.314 | 10.08 | | 10.09 | 9.235 | 11.13 |
| VG (DT) | 9.168 | 9.324 | 10.06 | | 10.09 | 9.240 | 11.06 |
| VG (DS) | 7.485 | 7.877 | 7.707 | | 6.716 | 6.185 | 6.53 |

| | Skewness | | | | Kurtosis | | |
|----------|----------|-------|-------|--|----------|-------|-------|
| Graph | T | P | B | | T | P | B |
| HVG | 1.724 | 1.530 | 1.628 | | 7.803 | 6.364 | 6.982 |
| HVG (DT) | 1.738 | 1.542 | 1.612 | | 7.926 | 6.439 | 6.824 |
| HVG (DS) | 1.766 | 1.642 | 1.723 | | 7.996 | 7.180 | 7.554 |
| VG | 2.920 | 2.788 | 2.602 | | 14.65 | 13.89 | 11.41 |
| VG (DT) | 2.918 | 2.786 | 2.597 | | 14.62 | 13.88 | 11.44 |
| VG (DS) | 4.070 | 2.950 | 3.659 | | 37.92 | 20.19 | 31.00 |

Table 7.2: First four moments of the degree for the HVG, VG and their detrended (DT) and deseasoned (DS) counterparts, computed over the full series of Turin (T), Prague (P) and Bologna (B).

In order to compare the irreversibility ratio IR, the series have been divided

| | Turin | Prague | Bologna |
|--------|--------|---------|---------|
| min | -7.8°C | -20.5°C | -8.2°C |
| max | 39.7°C | 37.8°C | 40.7°C |
| peak 1 | 9°C | 5°C | 8°C |
| peak 2 | 26°C | 20°C | 28°C |

Table 7.1: Minimum, maximum temperature and peaks of relative frequency for the daily maximum temperature series of Turin, Prague and Bologna.

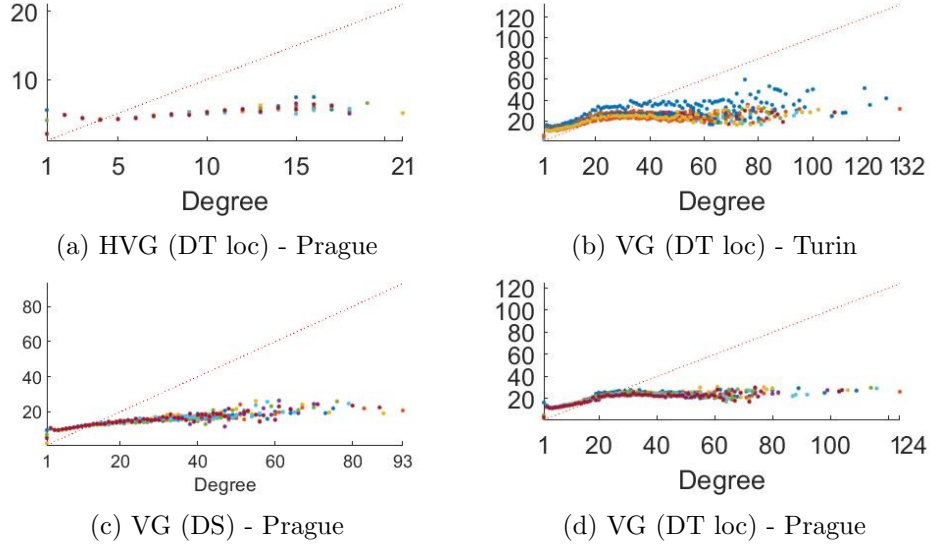


Figure 7.1: Examples of neighborhood connectivity plots for the locally detrended (DT loc) and deseasoned (DS) visibility graphs; in these cases the series of Prague and Turin are divided in 7 and 10 intervals respectively. The bisector $y = x$ is highlighted with a red dashed line for each plot.

in 21 and 43-year long intervals, starting from 1828 and ending in 2003. The results are presented in tables 7.3 and 7.4, and they show that the series of Prague and Bologna tend to be similar in terms of irreversibility. The deseasoned VGs generally feature lower values of IR with respect to the corresponding VGs, which leads to a higher number of reversible intervals with weak confidence; similarly, the detrended HVGs are characterized by a lower number of irreversible intervals compared to the corresponding HVGs. This applies to both detrended variants in the case of 21-year long intervals, but it only holds for the “globally” detrended HVGs if 43-year long intervals are considered.

As already mentioned, the relative frequency of motifs detected is quite similar for all case studies: for HVGs and deseasoned HVGs the most frequent motifs are A and D, followed by B, whereas for VGs and deseasoned VGs there is a much higher frequency of motif D, followed by motifs B and A. The motif frequency plots associated to the case study of Turin have been reported in Fig. 7.2 to provide a quantitative reference.

| Irreversibility ratio (IR) | | | | | | | | | | | | |
|----------------------------|-----|---|---|----|---|---|--------|---|---|----|---|---|
| | HVG | | | DS | | | DT loc | | | DT | | |
| Period | T | P | B | T | P | B | T | P | B | T | P | B |
| 1828-1849 | | | | | | | | | | | | |
| 1850-1871 | | | | | | | | | | | | |
| 1872-1893 | | | | | | | | | | | | |
| 1894-1915 | | | | | | | | | | | | |
| 1916-1937 | | | | | | | | | | | | |
| 1938-1959 | | | | | | | | | | | | |
| 1960-1981 | | | | | | | | | | | | |
| 1982-2003 | | | | | | | | | | | | |

| | VG | | | DS | | | DT loc | | | DT | | |
|-----------|----|---|---|----|---|---|--------|---|---|----|---|---|
| Period | T | P | B | T | P | B | T | P | B | T | P | B |
| 1828-1849 | | | | | | | | | | | | |
| 1850-1871 | | | | | | | | | | | | |
| 1872-1893 | | | | | | | | | | | | |
| 1894-1915 | | | | | | | | | | | | |
| 1916-1937 | | | | | | | | | | | | |
| 1938-1959 | | | | | | | | | | | | |
| 1960-1981 | | | | | | | | | | | | |
| 1982-2003 | | | | | | | | | | | | |

Table 7.3: Irreversibility ratio for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 8 intervals that are 21-years long. The coloring of the cells associates green to $IR \leq 1$ (reversibility), lime to $1 < IR \leq 4$ (reversibility with weak confidence), yellow to $4 < IR \leq 10$ (irreversibility with strong confidence) and orange to $IR > 10$ (irreversibility with extreme confidence).

| Irreversibility ratio (IR) | | | | | | | | | | | | |
|----------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| | HVG | | | DS | | | DT loc | | | DT | | |
| Period | T | P | B | T | P | B | T | P | B | T | P | B |
| 1828-1871 | Orange | Yellow | Orange | Yellow | Orange | Orange | Yellow | Orange | Orange | Green | Green | Yellow |
| 1872-1915 | Yellow | Green | Green | Green | Orange | Orange | Yellow | Orange | Orange | Green | Yellow | Green |
| 1916-1959 | Orange | Yellow | Green | Yellow | Orange | Orange | Yellow | Orange | Green | Yellow | Green | Green |
| 1960-2003 | Green | Yellow | Orange | Green | Yellow | Orange | Green | Orange | Orange | Green | Green | Yellow |

| | VG | | | DS | | | DT loc | | | DT | | |
|-----------|--------|--------|--------|-------|--------|--------|--------|--------|--------|--------|--------|--------|
| Period | T | P | B | T | P | B | T | P | B | T | P | B |
| 1828-1871 | Green | Orange | Orange | Green | Green | Yellow | Orange | Orange | Orange | Yellow | Orange | Orange |
| 1872-1915 | Green | Orange | Yellow | Green | Green | Green | Orange | Yellow | Yellow | Yellow | Yellow | Yellow |
| 1916-1959 | Yellow | Orange | Orange | Green | Yellow | Green | Orange | Yellow | Yellow | Orange | Yellow | Yellow |
| 1960-2003 | Green | Yellow | Orange | Green | Green | Green | Yellow | Yellow | Yellow | Green | Yellow | Orange |

Table 7.4: Irreversibility ratio for the HVG, VG and their detrended (DT/DT loc) and deseasoned (DS) counterparts, computed over 4 intervals 43-years long. The coloring of the cells associates green to $IR \leq 1$ (reversibility), lime to $1 < IR \leq 4$ (reversibility with weak confidence), yellow to $4 < IR \leq 10$ (irreversibility with strong confidence) and orange to $IR > 10$ (irreversibility with extreme confidence).

7.2 Final remarks

Three long term daily maximum temperature series have been analyzed: the series of Turin, Prague and Bologna span over 268, 180 and 190 years respectively. The homogenization of the first series is performed by taking into account the metadata, whereas the other two series have been homogenized by an automated procedure that is blind to metadata; in this way the climatic signal is isolated and can be considered as a reliable input for the following climate analysis. Data imputation is performed where necessary to eliminate a small and generally sparse number of missing values.

The HVG and VG algorithms are then employed to investigate the structural similarities of the underlying time series, and the same methods are also applied to the deseasoned and detrended series to get further insight; for each graph the first four moments of the degree, neighborhood connectivity, time reversibility and motif frequency are computed. As

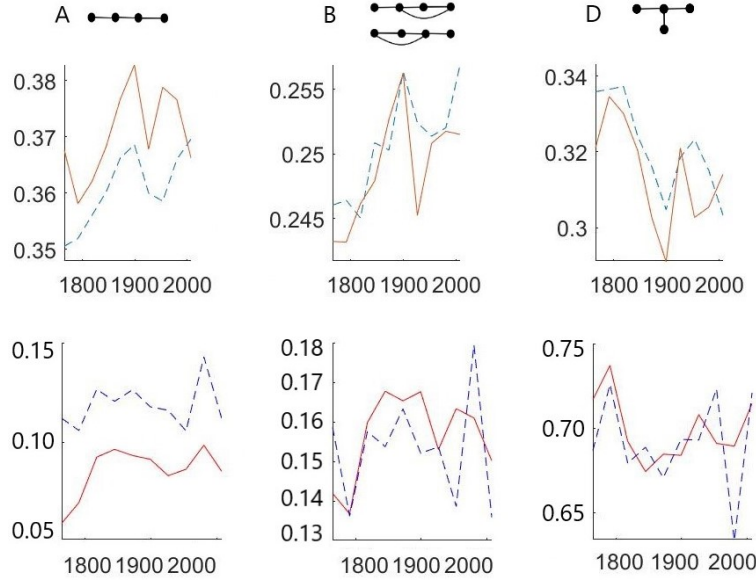


Figure 7.2: Frequencies of motifs A, B and D (left, center and right column respectively) for the HVGs (top row) and VGs (bottom row) computed over 10 intervals on the series of Turin. The orange and red lines correspond to the original underlying time series, and the blue dashed line to its deseasoned counterpart.

discussed in the previous section, the graphs have several qualitative and quantitative features in common: for instance they are disassortative, as high degree nodes tend to be linked to nodes of lower degree. In particular for the HVGs and their detrended and deseasoned counterparts the average degree of nodes connected to any given reference node in the graph does not depend on its degree.

The time irreversibility of subintervals of the series is evaluated through the irreversibility ratio: Lacasa et al. [14] introduced it in relation to the HVGs, but the same concept can be applied to the VGs. The results can be affected by the distinction between HVG and VG, as well as the choice of length of intervals. In particular the series evaluated through the VGs and their detrended counterparts may flag more intervals as irreversible with respect to the HVGs. The results that will be considered more relevant are those associated to the locally detrended and deseasoned VG; for these graphs there is a higher level of coherency between the two partitions considered, with most intervals being irreversible or at most reversible with

Irreversibility ratio (IR): VGs

| | DTDS | | | DS | | | DT loc | | |
|-----------|------|---|---|----|---|---|--------|---|---|
| Period | T | P | B | T | P | B | T | P | B |
| 1828-1849 | | | | | | | | | |
| 1850-1871 | | | | | | | | | |
| 1872-1893 | | | | | | | | | |
| 1894-1915 | | | | | | | | | |
| 1916-1937 | | | | | | | | | |
| 1938-1959 | | | | | | | | | |
| 1960-1981 | | | | | | | | | |
| 1982-2003 | | | | | | | | | |

Table 7.5: Irreversibility ratio for the locally detrended and deseasoned (DTDS), deseasoned (DS), and locally detrended (DT loc) VGs, computed over 8 intervals. The coloring of the cells associates green to $IR \leq 1$ (reversibility), lime to $1 < IR \leq 4$ (reversibility with weak confidence), yellow to $4 < IR \leq 10$ (irreversibility with strong confidence) and orange to $IR > 10$ (irreversibility with extreme confidence).

weak confidence. The choice of VGs over HVGs is justified by the ability of the former to retain more information about the structure of the underlying time series; similarly, decoupling the time series from its trend and/or seasonality allows an analysis of the irreversibility of the base signal, i.e. the oscillations in temperature. The presence of a local trend implicitly affects the computation of the average year of any given interval and is therefore accounted for to some extent, as shown in table 7.5; for this kind of graph the finer partition of the series may be preferable. The irreversibility ratio obtained for the deseasoned VGs tends to be lower than the corresponding values for the locally detrended VGs, confirming the impact of seasonality on the irreversibility of the underlying series.

The discrepancy of IR depending on the length of the intervals should be further investigated in order to justify the results obtained and provide a guideline for other similar applications of this method.

Bibliography

- [1] Baffo F., Desiato F., Lena F., Suatoni B., Toreti A., Bider M., Cacciamani C., Tinarelli G., Criteri di calcolo degli indicatori meteorologici, SCIA (Sistema Nazionale di raccolta, elaborazione e diffusione di dati Climatologici di Interesse Ambientale), (2005).
- [2] Brugnara Y., Auchmann R., Brönnimann S., Bozzo A., Berro D. C., Mercalli L., Trends of mean and extreme temperature indices since 1874 at low-elevation sites in the southern Alps, *Journal of Geophysical Research: Atmospheres* 2016.
- [3] Caussinus H., Mestre O., Detection and correction of artificial shifts in climate series, *Appl. Statist.* 53, Part 3, pp. 405-425 (2004).
- [4] Csörgő M., Horváth L., *Limit Theorems in Change-Point Analysis*, J. Wiley and Sons, 414 pp. (1997).
- [5] Chung-Ho W., Wen-Zer L., Hsiao-Chung T., Kuei-Yang W., The Sub-tropical Urban Island Effect Revealed in Eight Major Cities of Taiwan, WSEAS Int. Conf. on ENVIRONMENT, ECOSYSTEMS and DEVELOPMENT, Venice, Italy, November 2-4 (2005), pp14-20.
- [6] J.F. Donges, R.V. Donner, J. Kurths, Testing time series irreversibility using complex network methods, *Physical Review E*, 85, 046105p. (2012)
- [7] Firat M., Dikbas F., Gungor M., Analysis of temperature series: estimation of missing data and homogeneity test, *Meteorological Applications* 19, 397-406 (2012).
- [8] González-Espinoza A., Martínez-Mekler G., Lacasa L., Arrow of time across five centuries of classical music, *Physical Review Research* 2, 033166 (2020).

- [9] Iacobello G., Ridolfi L., Scarsoglio S., A review on turbulent and vortical flow analyses via complex networks, *Physica A* 563 (2021) 125476.
- [10] Iacovacci J., Lacasa L., Sequential visibility-graph motifs, *Physical Review E* 93, 042309 (2016).
- [11] Kehagias A., A hidden Markov model segmentation procedure for hydrological and environmental time series, *Stochastic Environ. Res. Risk Assess.*, 18, 117-130 (2004).
- [12] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, J. C. Nuno, From time series to complex networks: The visibility graph, *PNAS* (2008).
- [13] L. Lacasa, On the degree distribution of horizontal visibility graphs associated with Markov processes and dynamical systems: diagrammatic and variational approaches, *Nonlinearity* 27 (2014) 2063–2093.
- [14] Lacasa L., Nuñez A., Roldán E., Parrondo J.M.R., Luque B., Time series irreversibility: a visibility graph approach, *The European Physical Journal B* 85: 217 (2012).
- [15] Lyazhri F., Caussinus H., Choosing a Linear Model with a Random Number of Change-Points and Outliers, *Annals of the Institute of Statistical Mathematics* 49, 761-775 (1997).
- [16] Mateus C., Potito A., Curley M., Reconstruction of a long-term historical daily maximum and minimum air temperature network dataset for Ireland (1831-1968), *Geoscience Data Journal* (2020).
- [17] Di Napoli G., Mercalli L., *Il clima di Torino*, SMS (2008)
- [18] Mestre O., Méthodes statistiques pour l’homogénéisation de longues séries climatiques, *PhD Thesis* Université Paul Sabatier, Toulouse (2000).
- [19] Newman M. E. J., Mixing patterns in networks, *Physical Review E*, vol. 67, Issue 2, id. 026126, (2003).
- [20] Roldán E., Parrondo J. M. R., Entropy production and Kullback-Leibler divergence between stationary trajectories of discrete systems, *Physical Review E* 85, 031129 (2012).
- [21] Robinson P. J., On the Definition of a Heat Wave, *Journal of Applied Meteorology*, vol.40, p.762-774 (2001).

- [22] Roldán E., Barral J., Martin P., Parrondo J. M. R., Julicher F., Arrow of time in active fluctuations, arXiv:1803.04743.
- [23] Squintu A. A., van der Schrier G., Brugnara Y., Tank A. K., Homogenization of daily ECA&D temperature series, *International Journal of Climatology*, Vol. 39 Issue 3, p. 1243.1261 (2018)
- [24] Toreti A., Kuglitsch F.G., Xoplaki E., Luterbacher J., A Novel Approach for the Detection of Inhomogeneities Affecting Climate Time Series, *Journal of Applied Meteorology and Climatology*, vol.5, issue 2, 317-326 (2012).
- [25] Wang X. L., Wen Q. H., Wu Y., Penalized Maximal t Test for Detecting Undocumented Mean Change in Climate Data Series, *Journal of Applied Meteorology and Climatology*, vol. 46 issue 6, 916-931 (2007).
- [26] Welch L. R., Hidden Markov models and the Baum-Welch algorithm, *IEEE Inf. Theory Soc. Newsl.*, 53, 10-13 (2003).
- [27] World Meteorological Organization (WMO), Guide to Climatological Practices, 2018 edition.
- [28] Xu X., Zhang J., Small M., Superfamily phenomena and motifs of networks induced from time series, *Proceedings of the National Academy of Sciences*, (2009).
- [29] Zanchettin D., Pausata F.S.R., Khodri M., Timmreck C., Graf H., Claus J.H.J., Robock A., Rubino A., Thompson V., Toward predicting volcanically-forced decadal climate variability, *Past Global Changes*, 25(1), 25-31 (2017).
- [30] Zou Y., Donner R. V., Marwan N., Donges J. F., Kurths J., Complex network approaches to nonlinear time series analysis, *Physics Reports* 787 (2019) 1-97.
- [31] Iacobello G. (2021), Fast Horizontal Visibility Graph (HVG) for MATLAB, MATLAB Central File Exchange.
<https://www.mathworks.com/matlabcentral/fileexchange/72889-fast-horizontal-visibility-graph-hvg-for-matlab>
- [32] Iacobello G. (2021), Fast natural visibility graph (NVG) for MATLAB, MATLAB Central File Exchange.
<https://www.mathworks.com/matlabcentral/fileexchange/70432-fast-natural-visibility-graph-nvg-for-matlab>