Earth Science Informatics (2019) 12:685–697
https://doi.org/10.1007/s12145-019-00393-0

SOFTWARE ARTICLE

KarsTS: an R package for microclimate time series analysis

M. Sáez[1] · C. Pla[2] · S. Cuezva[3] · D. Benavente[1]

Received: 4 January 2019 / Accepted: 25 June 2019 / Published online: 6 July 2019

Abstract

KarsTS 2.2 is free, open-source, R-based software for microclimate time series, especially suited to the study of underground or highly insulated environments. The time series of interest include air temperature, humidity, and CO_2 and ^{222}Rn content, amongst others. These time series usually pose problems such as gaps, outliers, noise or relative shortness. KarsTS was born as a package for gap filling and thus, it offers multiple univariate and multivariate gap-filling tools well suited to these variables. However, as KarsTS was intended to be a self-sufficient program, it soon grew to encompass several tools for linear and nonlinear time series analysis, preprocessing and plotting. Indeed, many of these variables show a nonlinear behavior that is often disregarded; for this reason, we aim to spread and facilitate the use of some methodologically appropriate analysis tools, even amongst researcher that do not feel comfortable using a console. In this paper, we introduce an overview of KarsTS functionality and we show its potential through some practical application examples on four-year time series of temperature from the Rull cave (Spain).

Keywords Microclimate · Caves · R package · Nonlinear · Missing values · Recurrence analysis

Introduction

Microclimatic characterization is usually based on the continuous monitoring of several parameters as air temperature, relative humidity, pressure and gaseous carbon dioxide (CO_2) and radon (^{222}Rn) contents, amongst others (Cuezva et al. 2011). Microclimate characterization of such environments is a topic of great interest in the fields of cave art conservation (Bourges et al. 2014), historical heritage conservation (Camuffo et al. 2004), water resources management (Poulain et al. 2015), speleothems and paleoclimate reconstruction (Fairchild et al. 2006), human safety in underground environments (Alvarez-Gallego et al. 2015), and gases concentration and their interactions with the external atmosphere (Fernandez-Cortes et al. 2015), including contributions to the global carbon cycle and its role in climatic change (Garcia-Anton et al. 2017), amongst others.

Subsurface environments such as caves, basements or mines are, in general, isolated from the outside by layers of rock and soil. This results in thermally stable atmospheres, commonly saturated in water vapor and enriched in CO_2 and ^{222}Rn. The inner microclimate depends on the outside climate but also on other factors such as the existence and location of openings, the characteristics and thickness of the bedrock and the soil, etc. As a consequence, the measured time series and the relationships between them can be extremely complex (Perrier and Richon 2010; Baldini et al. 2006; Bourges et al. 2014).

Missing values are another major problem in environmental time series. Karst series are particularly affected by the existence of gaps. The singular environmental conditions and the presence of animals or vandals may damage the measuring devices, causing a wide variety of gaps in the registered data. Long gaps, which can last several days or even weeks, even when they are no numerous, are the most problematic.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12145-019-00393-0) contains supplementary material, which is available to authorized users.

✉ D. Benavente
david.benavente@ua.es

M. Sáez
marinasaez_andreu@hotmail.com

C. Pla
c.pla@ua.es

S. Cuezva
scuezva@ual.es

[1] Departamento de Ciencias de la Tierra y del Medio Ambiente, Universidad de Alicante, Alicante, Spain

[2] Departamento de Ingeniería Civil, Universidad de Alicante, Alicante, Spain

[3] Departamento de Biología y Geología, Universidad de Almería, Almería, Spain

KarsTS 2.2 is a multiplatform code with a friendly graphical interface for the analysis of microclimate time series. It is designed to address the specific set of problems that researchers face in the field of underground or insulated environments, such as caves or historical stone buildings; for this reason, amongst KarsTS broad functionality, its most noteworthy tools are related to gap filling and nonlinear analysis.

KarsTS is based on R, which is a cross-platform open-source computing environment, freely-available under the GNU General Public License (Grunsky 2002). R relies on a system of more than 10,000 contributed open-source packages, which enhance greatly its functionality. Several R-packages are devoted to missing values and nonlinear time series. The packages *mice* (Buuren and Groothuis-Oudshoorn 2011), *Amelia* (Honaker et al. 2011), *mi* (*Su* et al. 2011], *missForest* (Stekhoven 2013) and *Hmisc* (Harrell 2017) include tools for missing values imputation; however, many of their methods are not specific for time series and require normally distributed data. Regarding the packages for nonlinear time series, the package *GPoM* is devoted to nonlinear systems modelling (Mangiarotti et al. 2012a, 2012b). The package *nlts* provides tools for nonparametric autoregression and tests for linearity (Bjornstad 2017). The package *tseriesEntropy* offers an entropy measure and some tools derived from it (Giannerini 2017). The package tseriesChaos provides a number of analytic tools such as the correlation integral or the Lyapunov exponents (Di Narzo and Di Narzo 2013). Unfortunately, these methods generally require long time series, which are hardly available in microclimate research. Recurrence analysis, however, is a methodology well-suited for the analysis of nonlinear, short and noisy observational time series (Marwan et al. 2007). It experienced great development in the last decades (e.g., Bradley and Mantilla 2002; Marwan and Kurths 2005; Romano et al. 2005; March et al. 2005; Thiel et al. 2008; Marwan 2011) and it has been applied successfully to other disciplines, such as climatology (Marwan et al. 2003), materials science (Nichols et al. 2006), economy (Strozzi et al. 2007), physiology (Webber 2012) and seismology (Garcia et al. 2013). The package *nonlinearTseries* includes tools for recurrence plotting and quantitative recurrence analysis (QRA), although they are quite computationally expensive (Garcia 2015). The R package *fNonlinear* includes a function for plotting recurrence plots too, but not for QRA (Wuertz et al. 2017). The package *crqa* is devoted to cross-recurrence quantification analysis between two time-series of categorical or continuous values (Coco and Dale 2014). The fact that R and RStudio are manipulated mainly via code lines represents a barrier in many cases; because of this, R graphical interfaces are becoming more popular (for instance, RKWard, the Sciviews Virtual Box (Grosjean 2014) and the EPack Plugin, which provides RCommander with time series functionality (Fox and Bouchet-Valat 2017)).

Aiming to expand the use of R to the study of microclimate data, we developed KarsTS 2.2, which is cross-platform, free software, available on the Comprehensive R Archive Network under the GPL (>= 2) license. It offers functions to analyze, fill, plot and manipulate linear and nonlinear time series, even if they are short. These tools make KarsTS a self-sufficient program, where the researcher can perform the entire process, including preprocessing, filling and analysis. Regarding gap filling, KarsTS includes a handful of univariate and multivariate methods. Some of them have been adapted and implemented from gap-filling techniques developed in the field of Ecology for CO_2 flux time series (e.g., Falge et al. 2001a, b; Dengel et al. 2013; Moffat et al. 2007; Zhao and Huang 2015). Graphics are also an essential part of KarsTS. In general, R graphics lack interactivity; however, KarsTS offers some interactive plots, where the user can select elements and perform different actions (zoom, remove points, get coordinates, etc.). Finally, regarding recurrence analysis, our goal was to favor calculation speed and efficient memory usage because microclimate time series can produce very large recurrence matrices. We aim to raise awareness of the nonlinear nature of many microclimatic time series, as well as to promote the use of consistent methodological tools even amongst scientist that do not feel comfortable using a console.

In this paper, firstly we present an overview of KarsTS functionality. Then, we describe the data sets and outputs and the interface structure. Finally, in the application examples, we illustrate on observational cave time series the potential of some methods that have not been applied previously in this field.

Design and implementation

Development aims

KarsTS was born as an interface for filling gaps; however, soon it grew up to encompass a variety of tools for time series manipulation and analysis. Part of KarsTS functionality comes from contributed R packages, whilst other functions have been developed specifically for KarsTS. One of the main goals guiding KarsTS development was to make it self-sufficient; in other words, the user does not need other software to complement KarsTS functionality. For this reason, we included many tools for time series preprocessing and plotting. Even though, it can be used in combination with the R console, which is useful also for regular R users because KarsTS contains functions that are not in other packages.

KarsTS filling functionality is quite complete, both in terms of options and methods. The user can study the distribution of gaps (directly from time series, including time series with changing sampling frequency), study their nature (Little's test), evaluate filling methods on artificial gaps and, finally,

apply a filling method selectively on a subset of gaps. The program gathers a handful of univariate and multivariate methods that are effective on karst microclimate time series. Some filling methods were excluded either because they led to unsatisfactory results or because our time series did not meet their requirements. Similarly, we had to exclude many options for outlier detection; therefore we had to develop an interactive function for manual detection and a filter for points with anomalous slopes.

We propose a filling method based on twin recurrent points, therefore, KarsTS needs to be equipped with recurrence tools, however, the utility of recurrence analysis goes much further as we superficially show in the application examples. Our time series usually last few years, but their sampling frequency has to be quite high; this results in long time series in terms of number of points. Other R packages containing recurrence functions collapsed under such long time series, therefore KarsTS recurrence functions had to be implemented anew, considering carefully efficiency and memory usage.

Many researchers in the field of karst microclimate feel uncomfortable using command lines; on another hand, nonlinear tools such as recurrence analysis are not widely known in this field despite their potential. Hence, this is a double barrier that can be difficult to overcome. KarsTS, having a graphical interface, is intended to lower that barrier and raise consciousness of the nonlinearity of most microclimate time series. In addition, KarsTS covers the methods commonly used in karst microclimate research (statistics, correlations, etc.), which is expected to encourage the users. In every function, the user is allowed to choose as many relevant inputs as possible. This might require an initial learning effort but it fosters KarsTS capability and flexibility, as well as the user's awareness. Along these lines, we have provided very few optional, default inputs because they might encourage an irresponsible use of the software. KarsTS checks exhaustively the inputs entered by the user (class, range, compatibility, etc.) and throws explicative messages when they are inadequate. We have been meticulous to support the user and to avoid nonsensical results going undetected or collapsing the program.

Functionality and methods

In general, KarsTS functions can be divided in three layers. The core functions perform the calculations. Some of them have been developed for KarsTS whilst others come from other packages. Core functions are wrapped in a second layer of functions that verify exhaustively the inputs provided by the user and throw explicative messages when they are inadequate. All the previous functions can be accessed directly from the R console, even when the interface is closed. The third layer includes the functions corresponding to the graphical interface.

In the Supplementary material, table S.1 shows a complete list of KarsTS functionality and table S.2, the core functions inherited from other packages. Now we will describe briefly KarsTS functionality, highlighting the tools developed specifically for this program. We will also provide some theoretical background discretionarily.

KarsTS functionality is divided in five menus. Two menus are devoted entirely to data set manipulation (time series and gap sets, respectively). Actions related to file manipulation (load, save, export, etc.) and those that require only elementary mathematical procedures are also located in these menus (for example, resampling, scaling, rounding, cumulative sum etc.).

The user can select a set of gaps from a time series (diverse criteria are available) in order to apply later a filling technique only to that set of gaps (for example, spline interpolation only for gaps shorter than six missing values). Testing the suitability of the filling techniques to a particular time series is recommendable; to accomplish this, KarsTS allows the creation of artificial sets of gaps. The result can be tested visually or analytically, since KarsTS also contains a function to calculate the error between the imputed and the observed values.

The Analysis menu is devoted to analytical procedures that produce non-graphical results, that is, new data sets and tables. It includes statistics, loess seasonal decomposition and smoothing, principal component analysis and tests for normality, stationarity and linearity, as well as tools for the analysis of recurrence. Recurrence is the return of the system to the same state after some time and its analysis can be useful to characterize various types of regimes from low-dimensional linear deterministic to nonlinear and stochastic-like dynamics.

Recurrence analysis is based on the Theory of Dynamical Systems; therefore, the microclimate is conceived as a dynamical system, that is, a set of interrelated variables evolving through time. According to the Takens' Theorem (Takens 1981), systems dynamics can be reconstructed by embedding the observed variables available (see 3.2.2 for an example). This is a powerful tool, since underground systems are difficult to access and they involve variables that cannot be measured.

The fundamental tool for recurrence analysis are recurrence matrices, which are succinctly presented here (for a complete background, see Marwan et al. (2007). In section 3.2, we provide an example of creation and interpretation of a recurrence matrix (**RM**).

Let $\mathbf{x}_k$ be an embedded time series:

$$x_k \in R^m, k = 1, 2, \ldots N, \tag{1}$$

where m is the embedding dimension.

A **RM** can be defined as follows:

$$RM_{i,j}(\varepsilon) = \Theta\left(\varepsilon - \|x_i - x_j\|\right), i, j = 1, \ldots, N, \tag{2}$$

where ε is a threshold distance and Θ is the Heaviside function (Marwan et al. 2007). Thus, $\mathbf{RM}_{i,j}$ (ε) equals zero when the

distance between two points (which characterizes the state of the system in the phase space) is greater than the threshold ε and it equals one otherwise. In other words, ones in **RM** express recurrence and zeros express lack of recurrence. QRA can provide some useful information such as the self repeating-rate (which is usually called determinism in literature) the predictability of the system, the presence of successive alternating states (laminarity) or the time that the system remains trapped in a certain state (trapping time). Recurrence analysis can be applied also to cross and joint (multivariate) recurrence matrices.

Recurrence plots – located in the Plots menu – are the graphical counterpart of recurrence matrices. Values equal to one are plotted as points, whereas values equal to zero are not plotted. The points form patterns that give fast visual information about the dynamics of the system (Fig. 9).

The maximum size of R objects is rather limited (4Gb in the best case), especially when working with Windows. This poses a problem to recurrence analysis because matrices based on microclimatic time series tend to be very large. Indeed, we could not reuse recurrence tools from other packages, instead we had to design more efficient functions to create and store recurrence matrices, cross-recurrence matrices, joint recurrence matrices, distance matrices and their respective plots. The matrices are stored in a specific sparse format, which conditions the performing of the quantitative recurrence analysis (QRA), therefore tools for QRA from other packages cannot be used on KarsTS recurrence matrices. Following the same criterion of efficiency, KarsTS includes tools for estimating the recurrence rate, determinism or self-repeating rate, laminarity and recurrence probability.

The Plots menu also contains tools for plotting time series, phase portraits and distance matrices. Optionally, the user can customize the graphics to a great extent: colors, line width, point size, labels, pixel size etc. Some of the plots are interactive. Interactive plots provide KarsTS with essential functionality; besides zooming plot sections, the user can get point coordinates and graphically remove points. The latter tool is quite useful to eliminate outliers manually because automatic removal of outliers is seldom possible since our time series are often nonlinear and non-stationary. The outliers are usually due to malfunctions of the measuring devices and they can be detected visually with ease, though (see section 3.2.2 for an example).

The Plots menu contains other graphics, such as histograms, false nearest neighbors and tools for analyzing time series correlation (linear correlation, average mutual information and cross recurrence probability).

Finally, the fifth menu is devoted to filling methods. The upper row contains univariate methods. It includes different types of interpolation coming from the *zoo* and *stinepack* packages. Interpolation is a good choice for small gaps (smaller than the time series period), but it fails to reproduce periodical or quasi-periodical behavior. Aiming to expand the usefulness of interpolation to longer gaps, we have included an additional feature that allows to perform the interpolation taking on account the position of the value inside the period (we will refer to it as position-wise interpolation). For instance, let be a time series with measurements every 60 min and a periodicity of one day and a missing value corresponding to 12:00 h. If this option is selected, only the values measured at 12:00 h will be taken on account to perform the interpolation. This feature is very useful when the gaps are longer than half a period (in this example, gaps of 12 missing values or more).

The position-wise mean value (PwMV), specifically implemented for KarsTS, is inspired by the Mean Diurnal Variation, a gap-filling method used for Eddy CO_2 fluxes (Moffat et al. 2007). In the PwMV, a missing value is replaced by the mean of the values located in analogous positions in the periods surrounding the gap; the length of the period is defined by the user and, thus, not limited to daily variations. Along the same line, KarsTS includes position-wise interpolations. Gaps can also be filled by fitting an ARIMA model to the data. Optionally, KarsTS can suggest the ARIMA parameters using internally the *auto.arima* function, which eases significantly the process. The lower row offers multivariate filling methods, which are useful when other time series provide information for filling the incomplete one. These methods include ARIMAX models, generalized additive models and a random forest algorithm from the *missForest* package. Finally, the Twins method combines the Look Up Table methodology (used in Ecology, see Moffat et al. (2007)) with an original recurrence analysis approach. Let $\{X_{i,j}\}$ be a multivariate time series of length N with M variables $m_1,m_2,\ldots m_M$, where $X_{i,j}$ the i-th value of the variable m_j is missing. The LUT methodology consists of finding points where the non-missing variables $m_1,m_2,\ldots,m_{j-1},m_{j+1},\ldots,m_M$ take equal or very similar values to $(X_{i,1}, X_{i,2}, \ldots,X_{i,j-1}, X_{j+1}, \ldots, X_{i,M})$. We propose to define this similarity taken on account that these points are points of a dynamical system. Some methods to fill missing values using the Theory of Dynamical Systems have been proposed (Amritkar and Kumar 1995; Zhao et al. 2009); however, these methods are recursive and their errors grow very fast for observational time series. Our method is not recursive, since the missing values are filled considering the values of the non-missing variables. Moreover, it allows using embedded variables, thus taking advantage of the principle of embedding, which allows the reconstruction of dynamical systems when one or more variables are missing. In this filling method, we consider two points to be equivalent when they are twins. Twin points are points that produce identical columns in the recurrence matrix (this implies that they are so close in phase space that they share the same neighborhood of points).

Data sets and outputs

The user can handle three types of data sets: univariate time series, recurrence matrices and gap sets. Time series are data frames having the time in date format in the first column and the values in the second column. Gap sets and recurrence matrices are lists. A gap set is a collection of missing values, given by their positions in a time series; the list contains additional information about the original time series (name, length, start date, etc.). Recurrence matrices are sparse matrices containing only the positions of the ones, as well as additional information (embedding dimension, delay etc.). During the KarsTS session, data sets are located in the environment KTSEnv, which is accessible from the R console. This enables the combined use of KarsTS and R. KarsTS data sets have very specific structures and KarsTS will only recognize data sets that match exactly those structures; therefore, users must be careful when modifying them by means of the R-console. Despite this, these strict data set formats are desirable because they allow verifying the appropriateness of the inputs exhaustively.

KarsTS data sets can be imported from or exported to csv or txt files. Initially, the researcher's time series are usually in a csv or txt file; in contrast, gap sets and recurrence matrices are typically created with KarsTS. Saving the data to R files allows storing multiple data sets in the same file and eliminates any compatibility problem. KarsTS rejects incorrect files; for example, time series with dates in the wrong order. Nonetheless, time jumps and different sampling periods are accepted; KarsTS processes the time series in order that internally their sampling period is constant.

Interface structure

KarsTS interface consists of a welcome window, the main window and independent additional windows for plots and warning messages. The main window is divided in four parts (Fig. 1), which will be described briefly from the bottom to the top. The lower space contains four buttons which serve the following purposes: i) set the working directory; ii) create a txt file with the contents of the output window; iii) open a short help document (for further information, the user can check the KarsTS User's Guide) and iv) open a file with information about KarsTS current version.

Over the space containing these four buttons, the input panel (IP) lies on the left side and the output window (OW), on the right. Non-graphical results such as tables appear on the output window; the user can write, copy, paste and delete since the window is editable. As we have already mentioned, the OW contents can be saved to a txt file.

The uppermost row hosts five menu-buttons, namely, Time Series, Gap Sets, Analysis, Plots and Filling. As we mentioned in section 2.2, the first two menus are devoted to preprocessing; the third and fourth, to analysis and the last one, to missing data imputation. When a menu-button is pressed, the buttons corresponding to that menu are displayed in the rows below. Each menu-button has a distinctive color, which is shared by its buttons (although the buttons color is a tone lighter). Every function needs specific inputs, therefore when the user presses a button, the input panel changes accordingly.

Graphical outputs appear on new windows. From these windows the user can copy the plot to the clipboard or save it to a png or tiff file (Fig. 2a). In some cases, it is also possible to select a section of the plot and zoom it; the zoomed section appears on another window (Fig. 2b).

Illustrative examples

To illustrate the potential of KarsTS, in this section we provide examples of pre-processing, gap filling and recurrence analysis based on a data set gathered under real observational conditions in the Rull cave (Alicante, Spain) from 11 to 22–2012 to 10-28-2016. We will consider the outside temperature, which was recorded every 30 min with an independent data logger (Onset, Bourne, MA, USA), and the inside temperature, measured with a HygroClip S3 sensor (for further details about the monitoring, see Pla et al. 2017). The natural gaseous dynamic of this cave is characterized by two different stages. The temperature difference between the outside and the inside controls the ventilation processes, which cause the CO_2 and ^{222}Rn concentration to reach their lowest values in the coldest months. The cave temperature variations are negligible in comparison to the outside temperature ones. By contrast, in summer, ventilation is blocked and diffusion processes recharge the cave with soil-produced CO_2; the outside temperature is also an essential control for the soil CO_2 production and, of course, for the cave temperature itself (Pla et al. 2016a, b, 2017, Garcia-Anton et al. 2017). In summary, the outside temperature is the most important control for the cave microclimate.

Gap filling example

In order to show the application of some filling techniques, we select a fragment of the outside temperature (Fig.3). Then we create artificial gaps of varied sizes (button *Artificial random gaps*), as shown in Table 1. In all cases, the total number of missing values in the time series is 576 (10% of the fragment length).

For comparison, the time series are filled using all KarsTS univariate filling techniques: spline and Stinemann's interpolations, position-wise spline and Stinemann's interpolations, ARIMA model and position-wise mean-value. The ARIMA parameters are estimated automatically, being the result (2,0,0)x(1,1,0). For the position-wise mean value we use, in

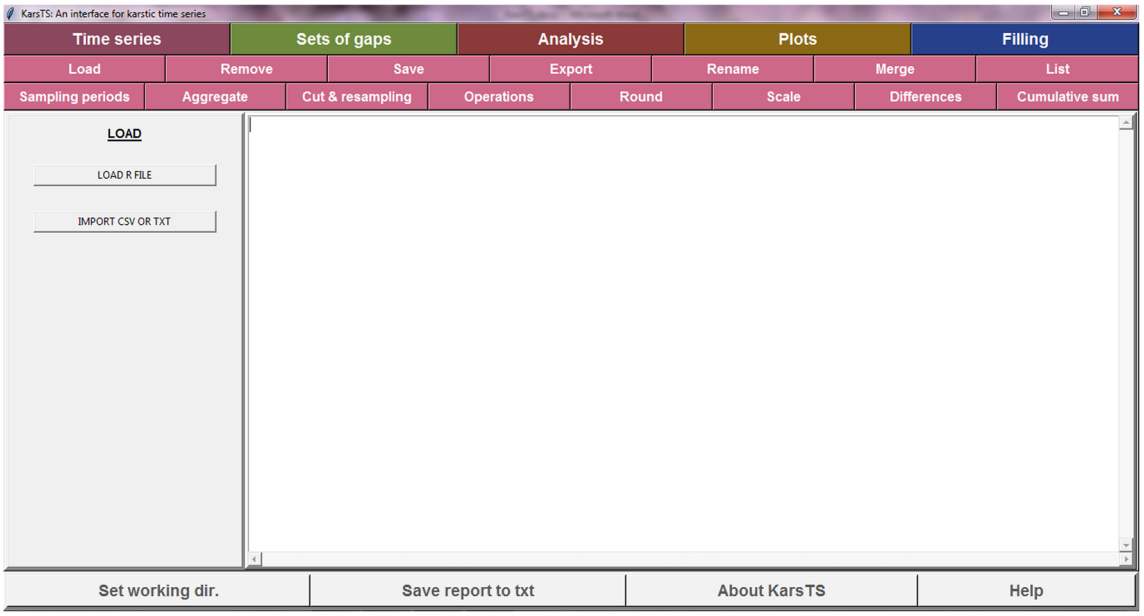

Fig. 1 KarsTS main window. At the top, the time series menu is displayed; at left, the input panel corresponding to the Load button and at right, the output window

general, a number of observations equal to twice the size of the gap; however, for gaps shorter than one period we take 96 observations (48 at right and 48 at left).

Table 2 shows the mean absolute error for each method (button *Check filling*). The interpolation error for the Spline interpolation and gaps of 12 h is 2.124 °C; however, the error grows fast as the length of the gaps is increased (for example, 35.770 °C for 6 days gaps). The Stinemann's interpolation error is somewhat smaller for 12 h gaps (1.948 °C) and its behavior for long gaps is slightly more stable. The position-wise interpolations (columns Pw-Splines and Pw-Stinemann) and the position-wise mean value (Pw-MV) entail a significant improvement. In these cases, the relationship between the gap length and the error magnitude is not systematical; however, the errors are of the same order for gaps up to 6 days (between 1.116 and 2.462 °C). The results for the ARIMA model are also of the same order, however, this method requires much more computation time. In conclusion, Table 2 shows that the position-wise Stinemann's interpolation offers the best results in this case. The position-wise spline interpolation would be a good choice too; indeed, we have sometimes observed a better behavior of the spline method over the Stinemann's one when the research requires to differentiate the time series (probably the spline interpolation is best to ensure smoothness).

Recurrence analysis of the temperature

In this example, we construct and interpret a **RM** from the time series of the difference between the outside and the inside temperature, which is the main control for ventilation.

Preprocessing

The sampling period is always 30 min and the first and last dates of both time series coincide. The time series of the outside temperature (Fig. 4a) has only nine missing values, therefore any filling method is suitable. However, the time series of the inside temperature (Fig. 5) has several long gaps. In addition, it contains many outliers, which are caused by malfunctions of the equipment. We clean the time series manually using the *Remove points* button (Fig. 5).

Next, we smooth the time series to remove noise and very high frequencies (button *Loess smooth.*). We have to choose the parameter alpha, which controls the degree of smoothing; in this case, alpha = 0.015 for both time series (the smoothed time series are not shown). For each point in the time series, the loess smoothing performs a local fitting in a neighborhood around the point; alpha indicates the length of this neighborhood as proportion to the total length of the time series (1435 days). Therefore, in this case the neighborhood includes 1435 *0.015 ≈ 21.5 days. This window is enough to eliminate noise and high frequency oscillations (of few days) from both the CO_2 and the ^{222}Rn time series. After smoothing, the cave temperature shorter periodicity is one year. Therefore we can now interpolate both time series with splines.

Finally, we subtract the inner temperature from the outside one (button *Operations*) in order to get the smoothed centered temperature (Fig. 4b, blue line). For the sake of simplicity, we rename the time series to cT.Sm, where Sm stands for *smooth*.

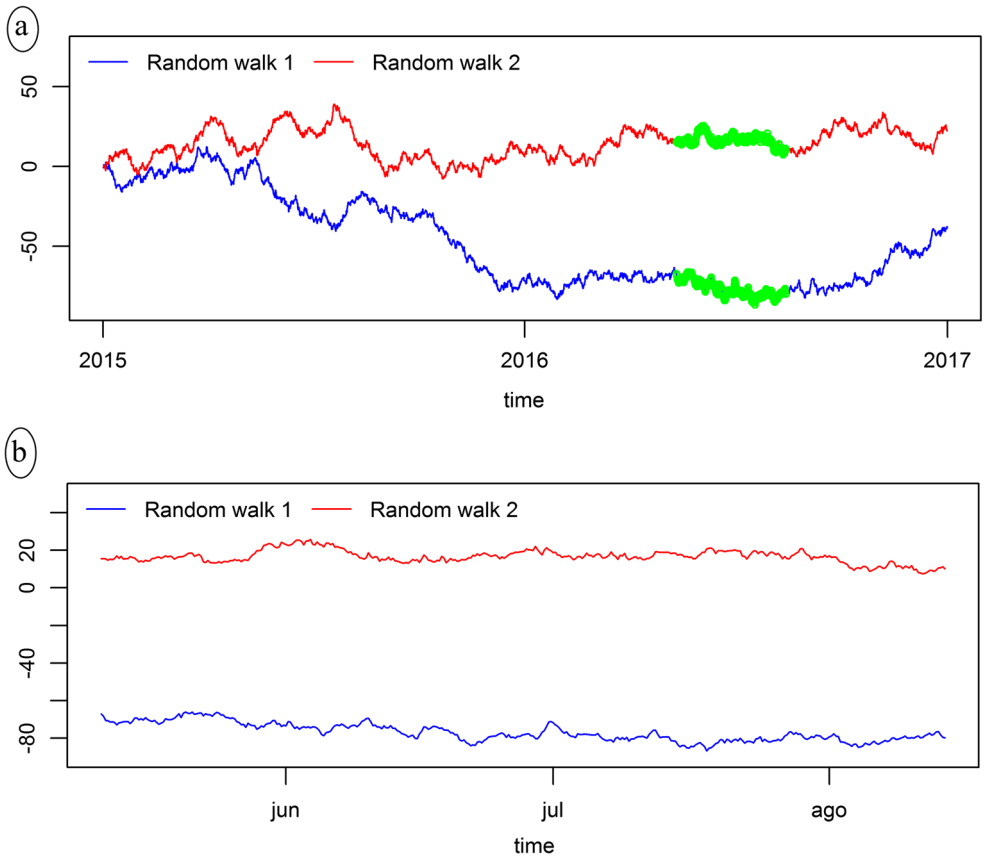

Fig. 2 a) Plot containing two time series currently under consideration; the section in green has been selected using the mouse. b) Zoom of a section of the time series. The zoomed sections can be turned into new time series using the button Create ts

cT.Sm is too long for a recurrence analysis: 68880 values. The corresponding **RM** would have 2,372,192,760 elements in each triangle. Even if KarsTS stores it as a sparse matrix, it would likely exceed R capacity of allocation. Therefore, we resample the time series (button *Cut and Resampling*) to get a time series with one value every 6 h: cT.Sm.6 h (not shown in any figure).

Estimation of the embedding parameters

Embedding is not strictly necessary to calculate recurrence matrices because the unembedded matrix contains, in principle, all the information (March et al. 2005); however, it is useful since it unfolds the dynamical information included in the original signal, and it facilitates the interpretation of the

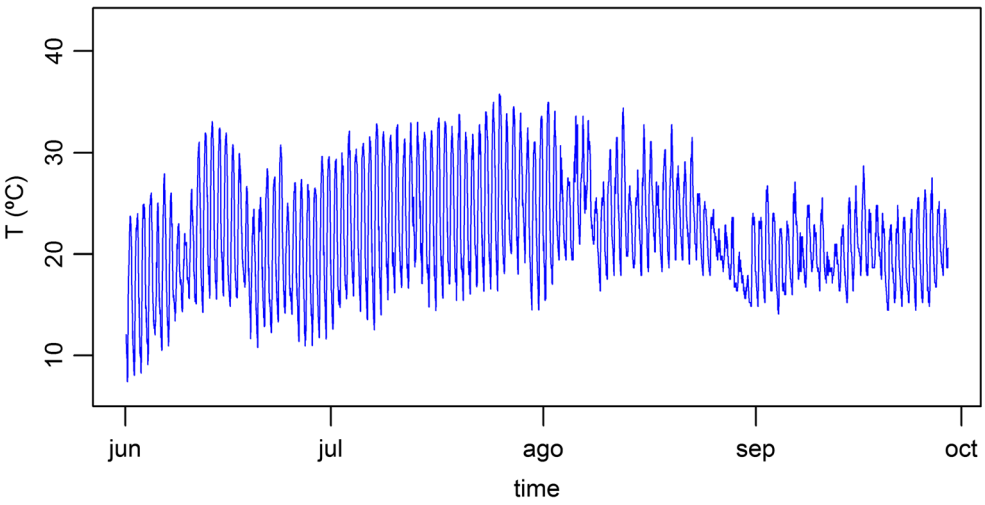

Fig. 3 Fragment of the outside temperature time series (°C) from June, 2013 to September, 2013

Table 1 set of time series with artificial gaps to test different filling methods. Note that the time series is sampled every 30 min, therefore every cycle (day) contains 48 observations

Time series name	Number of gaps	Missing values per gap	Gap length (days)
T.outside.1_24	24	24	0.5
T.outside.1_48	12	48	1
T.outside.1_96	6	96	2
T.outside.1_288	2	288	6
T.outside.1_576	1	576	12

recurrence plots. Therefore, choosing the embedding dimension (d), delay (τ) and threshold (ε) is a necessary step of recurrence analysis.

The most usual method to estimate the delay is to take the first minimum of the delayed average mutual information, AMI (Abarbanel et al. 1993) (button *Mutual*). It is also usual to take the first change of slope if there is no minimum. Note that embedding the time series causes a shortening equal to $(m - 1) \times \tau$; therefore, a distant minimum can lead to an unacceptable shortening in practice. Embedding also propagates gaps, which can be another drawback when m or τ are large.

Figure 6 shows the AMI plot for cT.Sm.6 h; clearly, the AMI has a minimum at lag = 360, that is, 90 days. However, this delay is probably too long considering the shortness of the time series. There is also a change of slope at 16 lags (4 days), approximately, which is likely a more feasible option. We have to consider, also, that the delay of 90 days will highlight the annual structures and the other will highlight the structures related to the intermediate time scales.

The next step is to estimate the embedding dimension, d. This is usually done using the false nearest neighbors method (Kennel et al. 1992) (button *FNN*). False neighbors are points in the embedding space that are close when the trajectory is compressed, but they separate when the system is conveniently unfolded, which happens when the embedding dimension is large enough. Figure 7 shows the FNN plot for the centered temperature time series. When the embedding dimension is around 5, the percent of false nearest neighbors becomes stable. For smaller dimension, the embedding is not ensured. For larger values, additional dimensions do not enable to unfold anymore information (see Letellier et al. (2008) for a discussion) and they might lead to spurious effects.

At this point, we have estimated the embedding dimension and the delay; now we have to estimate the threshold (ε). KarsTS offers functions to estimate it analytically (*Invariants* and *Ed(1) and Ed(2)*, see Supplementary material, table S.1), but they require longer time series. In this case, we will use distance plots (also known as unthresholded recurrence plots), which represent the distance between each pair of points by means of a color scale. We simply choose the threshold that visually produces clearer structures, in this case, short diagonal lines (Fig. 8). The user can apply the color scale to a range of distances; for example, we used progressively smaller distance ranges in order to refine the threshold: 0–15, 0–7 and 0–4 °C. From Fig. 8, we can estimate the threshold as 2 °C, although 1 °C seems to be a better choice in summer.

Distance plots provide good visualization, however, they cannot be used for QRA; for this reason, a RM is also needed (Fig. 9).

Temperature dynamics

Recurrence analysis of the centered temperature behavior is interpreted as a combination of deterministic chaos and laminarity at multiple scales. The large scale diagonal structures indicate that there is underlying determinism; however, rather than diagonal lines, they are succesions of square structures, which is a sign of laminarity. On the small scale, there are also diagonal and vertical structures that can be observed in Fig. 8 (distance plot) or Fig. 9 (recurrence plot). The plot forms a chess-like structure. In winter, there are large square structures filled with short diagonal lines (magenta in Fig. 8 and brown in Fig. 9, see example A). The fact that these structures are squared (not rhomboidal) evinces that the

Table 2 Mean absolute error (°C) of the observed versus imputed temperature values for different univariate filling techniques. Splines and Stinemann stand for Spline and Stinemann interpolations, respectively; Pw-Splines and Pw-Stinemann stand for position-wise spline and Stinemann's interpolation, Pw-MV stands for Position-wise mean value (see Suppl Mat 1,Table S.1)

Gap length (days)	Splines	Stinemann	Pw-Splines	Pw-Stinemann	Pw-MV	ARIMA
0.5	2.124	1.948	2.068	1.615	1.635	1.821
1	4.385	4.516	1.316	1.120	1.116	1.342
2	6.729	4.445	2.017	1.600	1.781	1.785
6	35.770	4.842	2.171	1.321	2.462	1.301
12	13.239	7.004	3.211	2.100	2.995	2.377

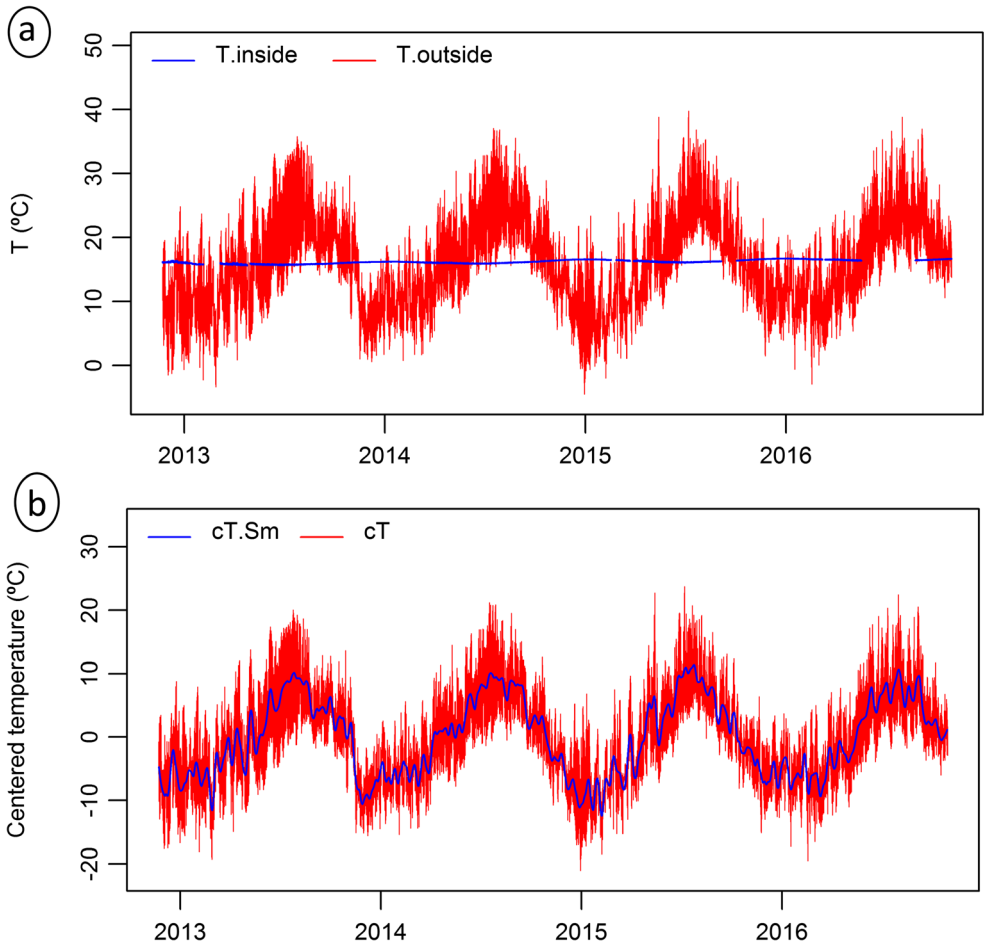

Fig. 4 **a** Outside temperature (red) and inside temperature (blue) time series (°C) **b** Centered temperature (red) and smoothed centered temperature (blue) time series (°C)

outside temperature surpasses and drops below the cave temperature very abruptly in comparison with the annual scale. The short diagonal lines inside correspond to temperature oscillations within the winter (Figs. 8 and 9, see example B). In summer, there are also squared structures filled with short diagonal lines, but they are smaller (magenta in Fig. 8 and brown in Fig. 9, example C). The diagonal lines in winter are more separated than they are in summer. This implies that the period of the temperature oscillations in winter is larger. As mentioned in section 3.3.2, the optimum threshold for the

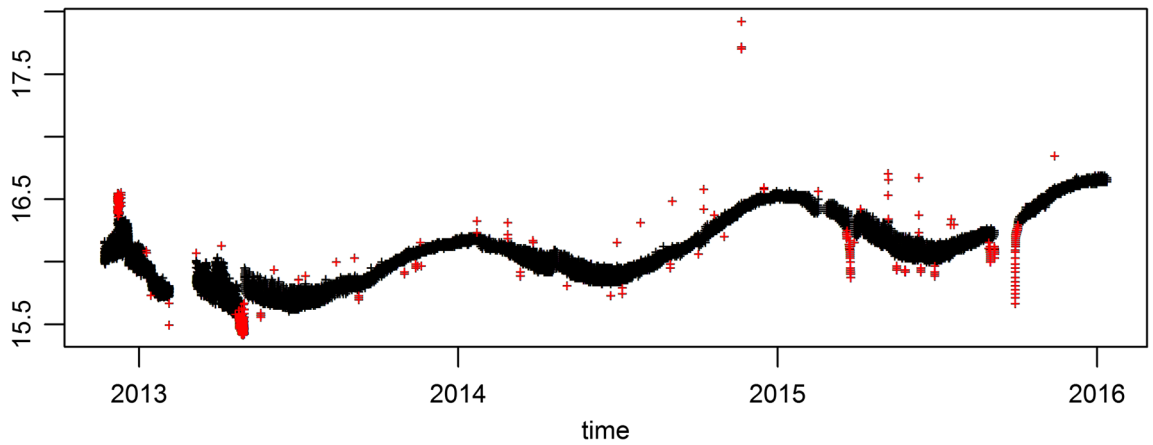

Fig. 5 Cave temperature (°C) before (in red) and after (in black) removing outliers with the Remove points button

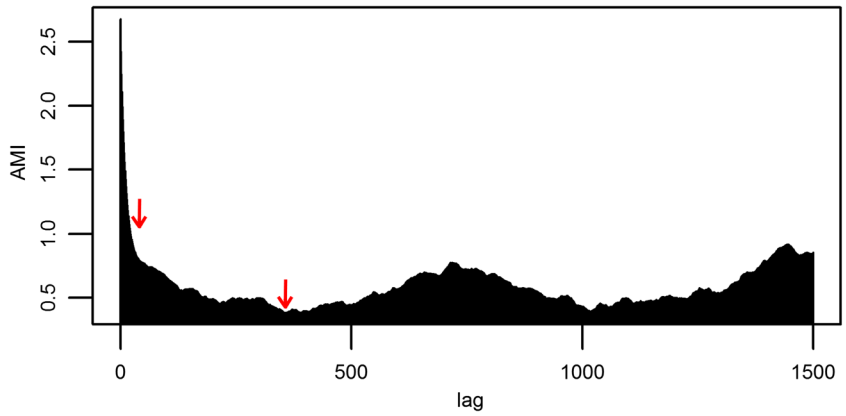

Fig. 6 Delayed average mutual information (AMI) as a function of delay time of the centered temperature (nats)

recurrence matrices seems to be 2 °C in winter and 1 °C in summer. Between the small summer and the large winter squares, there are more small squares. They correspond to sudden temperature changes followed by periods of stability. In other words, the temperature rises in few steps and reaches its maximum around July; then, it descends also in steps. After that, it enters the winter dynamics.

The outside temperature (Fig. 4a) and the centered temperature (Fig. 4b) time series have practically the same shape because the inner temperature is almost constant; as a consequence both temperatures would produce virtually identical recurrence matrices. The diagonal lines in the **RM** (Fig. 9) are discontinuous. This means that the centered temperature behaviour, although probably deterministic, cannot be predicted in the long term, which is characteristic of chaos.

We have estimated the recurrence rate, self-repeating rate and laminarity. The minimum self-repeating rate and laminarity were set to one day, which means that we consider the lines under this minimum length as noise.

The laminarity is very high in all cases (99%), which means that abrupt alternations are detected quite well. The self-repeating rate is (49%). This means that, approximately half of the times, the recurrence involves not only punctual states, but also the time evolution of the system during one day or more. The temperature predictability (mean of the recurrence diagonal lines lengths) is 8.46 days. Finally, the trapping time (mean of the recurrence vertical lines) is 7.67 days, which means that, on an average, the temperature remains stable for this time span before it changes.

The change between isolation (accumulation of CO_2 and ^{222}Rn) and ventilation is probably the most relevant factor when studying the gaseous dynamics of a cave. The Rull cave remains isolated from the outside's when the outside temperature is higher than the inner one (because the cave colder air remains trapped); in other words the change between isolation and ventilation must happen approximately when the centered temperature is close to zero or somewhat later. As we have shown in this paper, the outside temperature shows a strong degree of laminarity with stable states separated by abrupt temperature changes. This implies that the switch between isolation and ventilation is very sudden and therefore, it might be a delicate matter. Hence, we recommend studying the synchronization between gas concentration and temperature by means of recurrence matrices, which are able to deal with laminar and mixed behaviors. This is, however, out of the scope of this paper, where we have presented only one

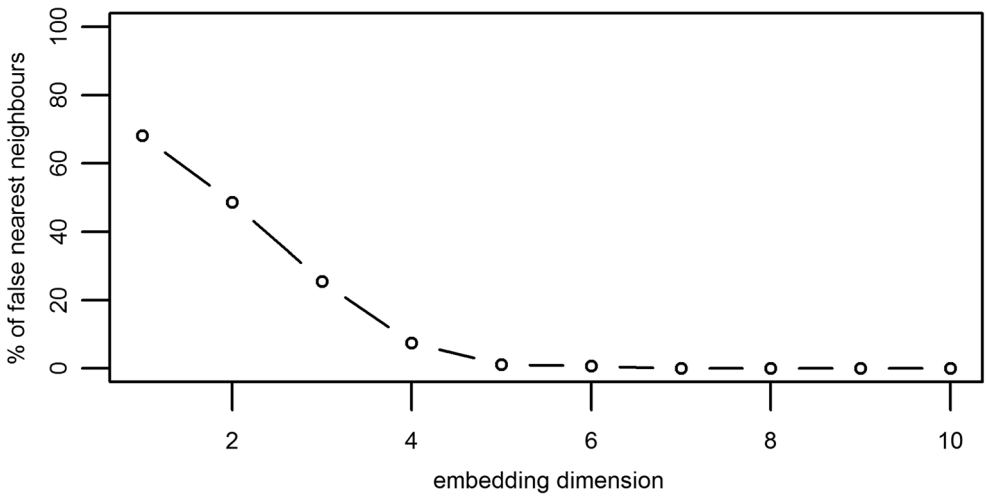

Fig. 7 False nearest neighbors plots for cT.Sm.6 h ($\tau = 16$)

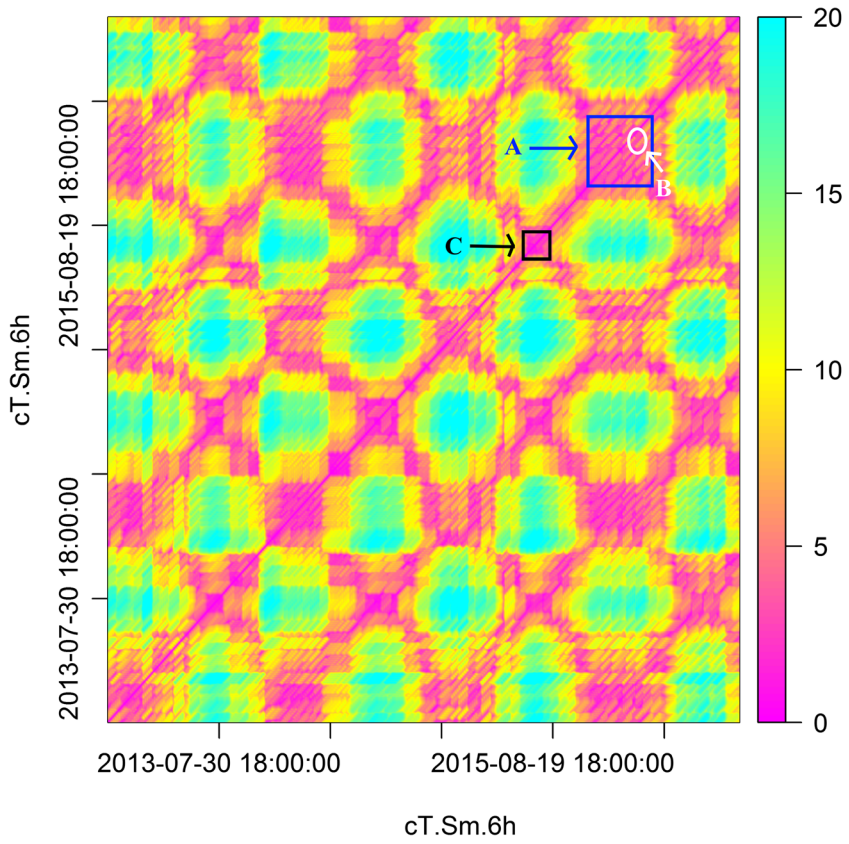

Fig. 8 Centered temperature distance plot (τ = 4 days; d = 5). A, B and C exemplify some essential structures (explained in section 3.2.3.)

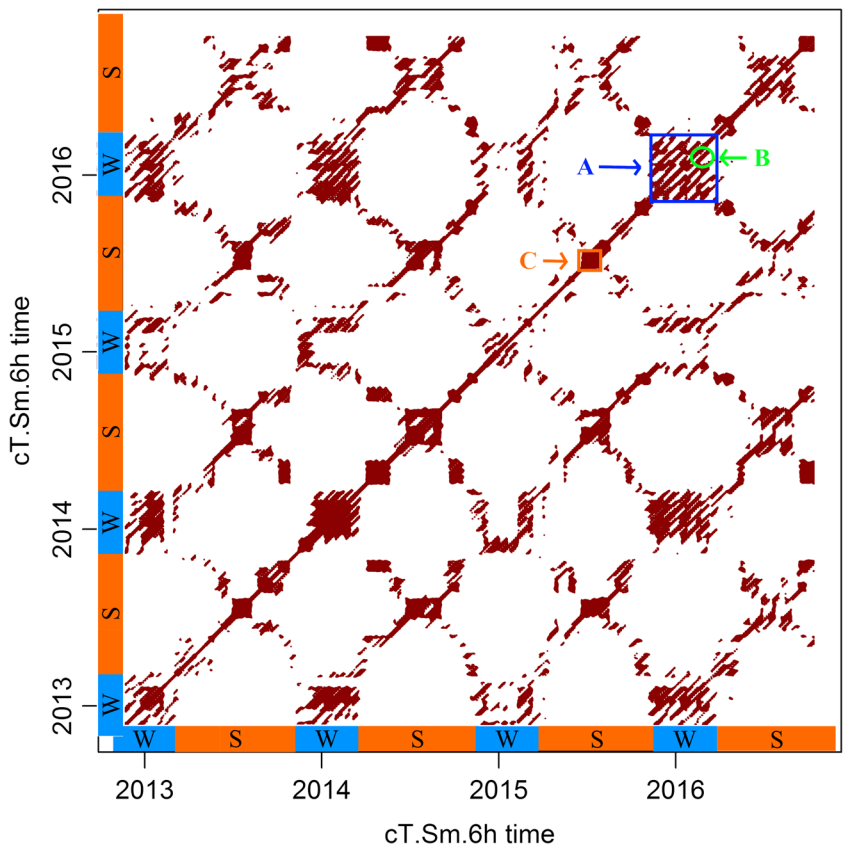

Fig. 9 Centered temperature RM (d = 5; τ = 4 days; r = 2 °C; Theiler's window = 2 days). A, B and C exemplify some essential structures (explained in section 3.2.3.)

variable, the outside temperature. This variable is the main driver for the gaseous interchange with the outside atmosphere (since the inner temperature is practically constant) and its effect over the cave gas concentration is rather fast. Therefore, the cave atmospheric composition is expected to show a similar behavior in terms of recurrence, predictability and trapping time.

Conclusions

We have developed KarsTS, an R-based, multiplatform package for microclimate time series with an emphasis on underground environments. It offers several tools for analysis, preprocessing and plotting, since we aspire to include everything the user needs to go through the entire characterization process. The functions and data sets can be handled via interface or via the R console.

Underground microclimate time series are the expression of real-world, complex, often nonlinear, processes, which is frequently overlooked. Hence, we aim to spread the use of appropriate methodological tools. KarsTS provides nonlinear tools including recurrence analysis since this technique is well adapted to observational time series, possibly nonlinear, noisy, short or incomplete. Recurrence analysis works on dynamical systems reconstructed by embedding; the possibility of reconstructing the system based on the variables available is also convenient because the monitoring of these environments is often difficult.

KarsTS also offers a handful of univariate and multivariate filling methods adapted from other fields of research and the possibility of applying them to different subsets of gaps within linear or nonlinear time series such as temperature and CO_2 or ^{222}Rn concentrations.

In this paper, we showed some application examples on four-year temperature time series from the Rull cave. On one hand, KarsTS univariate filling techniques were tested for different gap sizes; the better suited technique turned out to be the position-wise Stineman's interpolation. On another hand, we created and interpreted a **RM** based on the temperature difference between the outside and the inside, which is the main control for the cave microclimate and air composition. We found that the regime is strongly laminar and its predictability is approximately 8.5 days. The examples also included the pre-processing of the time series.

Acknowledgements This research was funded by the Spanish Ministry of Economy and Competitiveness Projects [CGL2011-25162, CGL2016-78318-C2-1-R, CGL2016-78318-C2-2-R and RTI2018-099052-B-I00]. A post-doctoral research fellowship was awarded to S. Cuezva by the University of Almería (Hipatia Programme). We also thank Dr. S. Mangiarotti for his useful discussions.

References

Abarbanel HD, Brown R, Sidorowich JJ, Tsimring LS (1993) The analysis of observed chaotic data in physical systems. Rev Mod Phys 65: 1331–1392

Alvarez-Gallego M, Garcia-Anton E, Fernandez-Cortes A, Cuezva S, Sanchez-Moral S (2015) High radon levels in subterranean environments: monitoring and technical criteria to ensure human safety (case of Castañar cave, Spain). J Environ Radioact 145:19–29

Amritkar R, Kumar PP (1995) Interpolation of missing data using nonlinear and chaotic system analysis. J Geophys Res-Atmos 100(D2): 3149–3154

Baldini JU, Baldini LM, McDermott F, Clipson N (2006) Carbon dioxide sources, sinks, and spatial variability in shallow temperate zone caves: evidence from Ballynamintra cave, Ireland. J Caves Karst Stud 68:4–11

Bjornstad ON (2017) nlts: (non)linear time series analysis. R package version 0.2–2

Bourges F, Genthon P, Genty D, Lorblanchet M, Mauduit E, D'Hulst D (2014) Conservation of prehistoric caves and stability of their inner climate: lessons from Chauvet and other French caves. Sci Total Environ 493:79–91

Bradley E, Mantilla R (2002) Recurrence plots and unstable periodic orbits. CHAOS 12:596–600

Buuren S, Groothuis-Oudshoorn K (2011) mice: multivariate imputation by chained equations in R. J Stat Softw 45

Camuffo D, Pagan E, Bernardi A, Becherini F (2004) The impact of heating, lighting and people in re-using historical buildings: a case study. J Cult Herit 5:409–416

Coco MI, Dale R (2014) Cross-recurrence quantification analysis of categorical and continuous time series: an R package. Front Psychol 5

Cuezva S, Fernandez-Cortes A, Benavente D, Serrano-Ortiz P, Kowalski A, Sanchez-Moral S (2011) Short-term CO_2(g) exchange between a shallow karstic cavity and the external atmosphere during summer: role of the surface soil layer. Atmos Environ 45:1418–1427

Dengel S, Zona D, Sachs T, Aurela M, Jammet M, Parmentier FJW, Oechel W, Vesala T (2013) Testing the applicability of neural networks as a gap-filling method using CH_4 flux data from high latitude wetlands. Biogeosciences 10:8185–8200

Di Narzo A, Di Narzo F (2013) tseriesChaos: Analysis of nonlinear time series. R package version 0.1–13

Fairchild IJ, Smith CL, Baker A, Fuller L, Spötl C, Mattey D, McDermott F (2006) Modification and preservation of environmental signals in speleothems. Earth Sci Rev 75:105–153

Falge E, Baldocchi D, Olson R, Anthoni P, Aubinet M, Bernhofer C, Burba G, Ceulemans R, Clement R, Dolman H, Granier A, Gross P, Grünwald T, Hollinger D, Jensen NO, Katul G, Keronen P, Kowalski A, Lai CT, Law BE, Meyers T, Moncrieff J, Moors E, Munger JW, Pilegaard K, Rannik Ü, Rebmann C, Suyker A, Tenhunen J, Tu K, Verma S, Vesala T, Wilson K, Wofsy S (2001a) Gap filling strategies for long term energy flux data sets. Agric For Meteorol 107:71–77

Falge E, Baldocchi D, Olson R, Anthoni P, Aubinet M, Bernhofer C, Burba G, Ceulemans R, Clement R, Dolman H, Granier A, Gross P, Grünwald T, Hollinger D, Jensen NO, Katul G, Keronen P, Kowalski A, Lai CT, Law BE, Meyers T, Moncrieff J, Moors E, Munger JW, Pilegaard K, Rannik Ü, Rebmann C, Suyker A, Tenhunen J, Tu K, Verma S, Vesala T, Wilson K, Wofsy S (2001b) Gap filling strategies for defensible annual sums of net ecosystem exchange. Agric For Meteorol 107:43–69

Fernandez-Cortes A, Cuezva S, Alvarez-Gallego M, Garcia-Anton E, Pla C, Benavente D, Jurado V, Saiz-Jimenez C, Sanchez-Moral S (2015) Subterranean atmospheres may act as daily methane sinks. Nat Commun 6:ncomms8003

Fox J, Bouchet-Valat M (2017) Rcmdr: R commander. R package version 2.4–1
Garcia CA (2015) nonlinearTseries: nonlinear time series analysis. R package version 0.2.3
Garcia SR, Romo MP, Figueroa-Nazuno J (2013) Characterization of ground motions using recurrence plots. Geof Inter 52:209–227
Garcia-Anton E, Cuezva S, Fernandez-Cortes A, Alvarez-Gallego M, Pla C, Benavente D, Cañaveras JC, Sanchez-Moral S (2017) Abiotic and seasonal control of soil-produced CO_2 efflux in karstic ecosystems located in oceanic and Mediterranean climates. Atmos Environ 164:31–49
Giannerini S (2017) tseriesEntropy: entropy based analysis and tests for time series. R package version 0.6–0
Grosjean P (2014) SciViews: a GUI API for R. UMONS Mons, Belgium
Grunsky EC (2002) R: a data analysis and statistical programming environment–an emerging tool for the geosciences. Comput Geosci 28:1219–1222
Harrell FE (2017) Hmisc: Harrell Miscellaneous. R package version 4.0–3
Honaker J, King G, Blackwell M (2011) Amelia II: a program for missing data. J Stat Softw 45:1–47
Kennel MB, Brown R, Abarbanel HD (1992) Determining embedding dimension for phase-space reconstruction using a geometrical construction. Phys Rev A 45:3403–3411
Letellier C, Moroz I, Gilmore R (2008) Comparison of tests for embeddings. Phys Rev E 78:026203
Mangiarotti S, Coudret R, Drapeau L, Jarlan L (2012a) Polynomial search and global modeling: two algorithms for modeling chaos. Phys Rev E 86:046205
Mangiarotti S, Mazzega P, Hiernaux P, Mougin E (2012b) Predictability of vegetation cycles over the semi-arid region of Gourma (Mali) from forecasts of AVHRR-NDVI signals. Remote Sens Environ 123:246–257
March TK, Chapman SC, Dendy RO (2005) Recurrence plot statistics and the effect of embedding. Physica D 200:171–184
Marwan N (2011) How to avoid potential pitfalls in recurrence plot based data analysis. Int J Bifurcat Chaos 21:1003–1017
Marwan N, Kurths J (2005) Line structures in recurrence plots. Phys Lett A 336:349–357
Marwan N, Trauth MH, Vuille M, Kurths J (2003) Comparing modern and Pleistocene ENSO-like influences in NW Argentina using nonlinear time series analysis methods. Clim Dyn 21:317–326
Marwan N, Romano MC, Thiel M, Kurths J (2007) Recurrence plots for the analysis of complex systems. Phys Rep 438:237–329
Moffat AM, Papale D, Reichstein M, Hollinger DY, Richardson AD, Barr AG, Beckstein C, Braswell BH, Churkina G, Desai AR, Falge E, Gove JH, Heimann M, Hui D, Jarvis AJ, Kattge J, Noormets A, Stauch VJ (2007) Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agric For Meteorol 147:209–232
Nichols JM, Trickey ST, Seaver M (2006) Damage detection using multivariate recurrence quantification analysis. Mech Syst Signal Process 20:421–437
Perrier F, Richon P (2010) Spatiotemporal variation of radon and carbon dioxide concentrations in an underground quarry: coupled processes of natural ventilation, barometric pumping and internal mixing. J Environ Radioact 101(4):279–296
Pla C, Cuezva S, Garcia-Anton E, Fernandez-Cortes Á, Cañaveras JC, Sanchez-Moral S, Benavente D (2016a) Changes in the CO2 dynamics in near-surface cavities under a future warming scenario: factors and evidence from the field and experimental findings. Sci Total Environ 565(565):1151–1164
Pla C, Galiana-Merino JJ, Cuezva S, Fernandez-Cortes Á, Cañaveras JC, Benavente D (2016b) Assessment of CO2 dynamics in subsurface atmospheres using the wavelet approach: from cavity–atmosphere exchange to anthropogenic impacts in Rull cave (Vall d0Ebo, Spain). Environ Earth Sci 75(6). https://doi.org/10.1007/s12665-016-5325-y
Pla C, Cuezva S, Martínez-Martínez J, Fernandez-Cortes Á, García-Antón E, Fusi N, Crosta GB, Cuevas-González J, Cañaveras JC, Sanchez-Moral S, Benavente D (2017) Role of soil pore structure in water infiltration and CO2 exchange between the atmosphere and underground air in the vadose zone: a combined laboratory and field approach. Catena 149:402–416
Poulain A, Rochez G, Bonniver I, Hallet V (2015) Stalactite drip-water monitoring and tracer tests approach to assess hydrogeologic behavior of karst vadose zone: case study of Han-Sur-Lesse (Belgium). Environ Earth Sci 74:7685–7697
Romano MC, Thiel M, Kurths J, Kiss IZ, Hudson JL (2005) Detection of synchronization for non-phase-coherent and non-stationary data. Europhys Lett 71:466–472
Stekhoven DJ (2013) missForest: nonparametric missing value imputation using random Forest. R package version 1.4
Strozzi F, Gutierrez E, Noe C, Rossi T, Serati M, Zaldivar JM (2007) Application of non-linear time series analysis techniques to the nordic spot electricity market data. Liuc Papers
Su YS, Gelman A, Hill J, Yajima M (2011) Multiple imputation with diagnostics (mi) in R: opening windows into the black box. J Stat Softw 45:1–31
Takens F (1981) Detecting strange attractors in turbulence. In: Rand D., Young LS. (eds) Dynamical systems and turbulence, Warwick 1980. Lecture notes in mathematics, vol 898. Springer, 366–381
Thiel M, Romano MC, Kurths J, Rolfs M, Kliegl R (2008) Generating surrogates from recurrences. Philos Trans Royal Soc A 366:545–557
Webber CL (2012) Recurrence quantification of fractal structures. Front Physiol 3
Wuertz D, Setz T, Chalabi Y (2017) fNonlinear: Rmetrics - nonlinear and chaotic time series modelling. R package version 3042.79
Zhao X, Huang Y (2015) A comparison of three gap filling techniques for eddy covariance net carbon fluxes in short vegetation ecosystems. Adv Meteorol 260580:12
Zhao P, Xingb L, Yuc J (2009) Chaotic time series prediction: from one to another. Phys Lett A 373:2174–2177

Availability and Requirements

Program title: KarsTS 2.2

Developer: Marina Sáez (email: marinasaez_andreu@hotmail.com)

Available from: https://cran.r-project.org/web/packages/KarsTS/index.html

Licensing provisions: GNU General Public License 2

Programming language: R (>= 3.4.0)

Software: minimum, Windows 7 or Mac OS v.10.11 . R (>= 3.4.0) and R Studio (>= 1.1.383).

Running time: Interactive

Software Files

Program title: KarsTS 2.2

Available from: https://cran.r-project.org/web/packages/KarsTS/index.html