Modelling Non-Equidistant Time Series Using Spline Interpolation*

Gábor Rappai
professor
University of Pécs
E-mail: rappai@ktk.pte.hu

This study gives an overview of spline interpolation, a special class of interpolation methods. The focal concern discussed in this paper is that the augmentation of non-equidistant time series (using averages, previous values, or interpolation) often leads to misleading or erroneous conclusions, as the augmented time series may have different characteristics than the original data generating process. The author's main purpose is to demonstrate that augmentations of any kind are to be planned carefully. To underline this statement, he applies the most frequently used methods on empirical time series, then collects and highlights the most prevalent conclusions.

Keywords:
Time series analysis.
Interpolation.

* The support and valuable advice the author has received from his colleagues and from the anonymous reviewer of the *Hungarian Statistical Review* are greatly appreciated.

The vast majority of literature addressing time series analysis focuses on uniformly spaced events, where the gaps between observations are of equal lengths. In such cases, the specific times or dates of the observations do not need to be recorded, as we can retain all information by only indicating the time of the first and last observations, along with the frequency of occurrences. These data sets are often referred to as equidistant time series, where the times or dates of observations are typically replaced with consecutive integers ($t = 1, 2, \ldots, T$) without losing any information.

However, time series analysis is not limited to such "ideal" cases, as the lengths of gaps between observations may vary. Financial markets, among many other fields within economics, are a great example: price charts often include gaps of different lengths, due to the presence of weekends and national holidays, but it is even more prevalent on intra-day levels as nothing guarantees that market transactions would take place at a regular pace, say, every 10 or 20 seconds. In macro-level demand models, for instance, non-observable (latent or induced) demand is often replaced with unequally spaced purchases.

Marine biology, astrophysics and meteorology apply a wide variety of models in order to overcome the challenges posed by non-equidistant observations. The reasons why such unstructured time series are captured would be worth a study by itself, including the measurement techniques and corresponding strategic considerations that may improve the results.[1] Even though this unstructuredness can be described as an anomaly, we treat this phenomenon as a given, including the difficulties and challenges that are involved, and that today's analysts have to face.

In attempt to overcome these difficulties, a handful of techniques and approaches have been developed over time.

1. In financial time series models (or even in meteorogical models addressing levels of precipitation), each gap between actual observations is usually filled with either zeros or with the value of the preceding observation.[2] Models, then, are based upon this "mended," augmented time series. This method, quite obviously, requires careful considerations as substitutions in large quantities can result in misleading conclusions.

[1] Three main categories can be differentiated within the group of unstructured time series, all of which are referred to as non-equidistant time series in this paper. These categories include structured but unequally spaced time series; structured time series with occasionally missing observations; and purely unstructured time series.

[2] Yield projection is one of the most frequent tasks in price modelling. In this case, replacing missing yields with zeros is equivalent to substituting the missing prices with the last preceding values (forward-flat interpolation).

2. Finding a frequency that matches every observation appears to be a more reasonable approach, as we can interpolate the observed values in order to substitute the missing values. The application of linear interpolations is undoubtedly a quick and convenient solution, although non-linearity tests often become less effective as they tend to indicate erroneously non-linearity (*Schmitz* [2000]).

3. A substantially more complex solution that this paper will not address in detail is creating an estimating function based on the covariance structure (autocovariance function) of the time series in order to fill in the missing values. In case a gap can be described with random variables having a probability distribution that is identical with that of the observed values, the Lomb-Scargle algorithm can be applied (see *Lomb* [1976] or *Schmitz* [2000]). This algorithm provides a periodogram of the missing values, and makes assumptions regarding gaps unnecessary.

4. The application of continuous-time models may be considered when the data series are genuinely unstructured. This issue has been addressed relatively early in modelling literature (see *Jones* [1985], *Bergstrom* [1985], or *Hansen–Sargent* [1991]). We also recommend the works of *Brockwell* [2001] or *Cochrane* [2012] who gave excellent summaries on this subject). For an exhaustive description of the state-space model based solution of parameter estimation, please refer to *Wang* [2013]. This model class is primarily designed for predictive purposes, however, the methods are based on the assumption that the time series, in fact, are equidistant, which can certainly be considered a weakness.

In this study, we focus on spline interpolation, a special class of interpolation methods. The focal issue being addressed is that the augmentation of gaps within non-equidistant time series (using averages, preceding values, etc.) often leads to incorrect and misleading conclusions as the augmented time series, after the augmentation, are treated as if they were, in fact, equidistant. Since the modelling techniques thereafter are based upon the assumption of equidistance, the original data generating function may have different characteristics than the ones that are implied based on the augmented time series. Another main purpose of this study is to illustrate that despite the fact that statistical software packages offer a variety of augmentation techniques "on a silver plate", they should not be routinely, "blindsightedly" applied.

After a brief overview of the most popular interpolation techniques, we are going to pay close attention to the basic features of spline interpolation. Following the section dedicated to this area, we address the risk of misspecifying data generating processes when the usual tests are run on augmented time series. To achieve this goal, we used computer-based simulations. In the final section of the paper, however, we

use empirical non-equidistant time series in order to better portray our conclusions and recommendations.

1. Reassumption of missing values

The typical way of handling non-equidistant time series is by assuming that there is, in fact, an original equidistant time series from which certain values are missing. The most frequently used step, then, is the application of time series interpolation.[3] Here, we discuss two popular types of interpolation:

– the linear type (or log-linear, as the concept of the two are very similar) and
– the spline approach.

Both types of interpolation begin with the same step: we have to determine the frequency that best describes the time series in order to identify the locations of missing values to be augmented. In certain cases, this step doesn't require careful considerations, as a "natural" frequency may be trivial, even though not all expected observations are recorded (e.g. the gaps in daily stock market closing prices where the natural frequency is one per day and every sixth and seventh value is missing). In many other situations, however, there is no such trivial frequency: this is the case, for example, in time series compiled from world records in sports or when we consider the proportions of mandates representing the power of a government coalition, where the gaps between occurrences are not supposed to be equally spaced in the first place. Regarding the stock markets, national holidays across countries may cause similar problems and lead to unstructuredness.

The general recommendation is to derive the hypothetical frequency from the smallest gap between the actual observations. This, however, does not necessarily result in a frequency to which all actual values can be fitted. When it comes to determining hypothetical frequencies, the core dilemma is the following:

– On one hand, most or all original observations should match the theoretical (augmented) time series, which is an argument for assuming high hypothetical frequency.

[3] Note that interpolation is a generic term used in many contexts: its use is not limited to augmenting gaps in time series, but every estimation technique is referred to as interpolation when an ex post estimate is used between a pair of observations.

– On the other hand, the number of values obtained with interpolation should not exceed (or, according to some arguments, get close to) the number of actual values. This is an argument against assuming excessively high hypothetical frequencies.[4]

Once the hypothetical frequency is determined and the time series that still includes missing values is converted to equidistant sets, interpolation means the estimation of missing values between each pair of known empirical values.

The augmented time series obtained after the interpolation should have two main characteristics (prerequisites). Preferably, it should:

– not differ from the original values where empirical observations exist, and
– be relatively smooth.

Let the empirical time series be

$$y_{t_1}, y_{t_2}, \ldots, y_{t_k}, \ldots, y_{t_T}$$

where the distance $t_2 - t_1$ isn't necessarily equal to the $t_3 - t_2$ distance, etc. Let Δ denote the largest distance and let each $t_k - t_{k-1}$ be equal to or divisible by Δ.[5]

From this, the following, incomplete time series can be constructed:

$$y_{t_1}, y_{t_1+\Delta}, y_{t_1+2\Delta}, \ldots, y_{t_1+j\times\Delta}, \ldots, y_{t_T}$$

where $y_{t_1+j\times\Delta}$ are values to be determined by interpolation when a $t_1 + j\times\Delta$ does not match any of the empirical t_k observations.

Essentially, the purpose of interpolations is to give an estimation of values corresponding to points of time where no empirical observations were made or recorded. Linear interpolation is a simple method that does not completely fulfil our previous prerequisites. If

$$y_{t_{k-1}} = y_{t_1+(j-1)\times\Delta} < y_{t_1+j\times\Delta} < y_{t_k} = y_{t_1+(j+1)\times\Delta}$$

[4] In the last two decades, statistical literature has placed substantial emphasis on the fact that these criteria are difficult to meet in the case of ultra-high frequency data sets. For an overview of methodological tools and techniques that can be applied in similar situations (such as analyses based on stock market transactions), please refer to *Engle* [1996].

[5] This is often but not necessarily the same as the smallest distance between any two actual observations.

where $y_{t_1+j\times\Delta}$ are originally missing but both of their neighbours are known, the supplemental values can be obtained using

$$\hat{y}_{t_1+j\times\Delta} = y_{t_{k-1}} + \frac{y_{t_k} - y_{t_{k-1}}}{t_k - t_{k-1}} = y_{t_{k-1}} + \frac{y_{t_k} - y_{t_{k-1}}}{2\Delta}.$$

In case there are multiple missing values between two observations, interpolation can be performed by adjusting the denominator. In general, linear interpolation can be formalized as

$$\hat{y}^{LIN} = (1-\lambda)\, y_{t_{k-1}} + \lambda y_{t_k} \qquad /1/$$

where $y_{t_{k-1}}$ is the last non-missing value, y_{t_k} is the subsequent non-missing value, and λ denotes the relative position of the missing value between two known empirical values. (If there is one missing value, the known difference has to be divided by two; in the case of two missing values, the difference is to be divided by three, etc.)

Linear interpolation, however, tends to result in hectic, "fractured" diagrams, i.e. this method often leads to estimating functions with wildly oscillating slopes. To "tame" this phenomenon, we often turn to log-linear interpolation where estimated values can be generated using

$$\hat{y}^{LOGLIN} = e^{(1-\lambda)\ln y_{t_{k-1}} + \lambda \ln y_{t_k}}. \qquad /2/$$

Even though the logarithmic function leads to smoother diagrams, other problems arise, as negative values cannot be directly dealt with. Therefore, despite their simplicity, linear and log-linear interpolations are typically recommended as exploratory tools only.[6]

Spline interpolation, as well, has been developed with the purpose of obtaining smooth functions.[7] According to the original definition of this method, an $S(t)$ interpolating function is assigned to each section, each of which has to satisfy certain conditions. Using the same notations as earlier, where $t_1 < t_2 < \ldots < t_T$ are the points of observations and assuming that the corresponding values depend on time,

[6] These methods can be further improved by considering additional values besides the ones immediately before and after a given missing value, which can result in smoother functions. One of these methods is the cardinal spline method, a built-in tool included in the Eviews software package.

[7] For the first mathematical reference to splines, see *Schoenberg* [1946].

i.e. $y_{t_k} = f(t_k)$, the goal is to find an $S(t)$ function that fulfils the following criteria:[8]

$$S(t) = S_{t_k}(t) \quad t \in [t_1, t_T], \qquad /C1/$$

$$S(t_k) = y_{t_k}, \qquad /C2/$$

$$S_{t_k}(t_{k+1}) = S_{t_{k+1}}(t_{k+1}). \qquad /C3/$$

These conditions, in a less formalized language, mean that interpolation can be performed piecewise, where each segment can be defined with a different function /C1/; the interpolating function's values match the original observations where they exist /C2/; and the curve that is obtained as a result of interpolation is continuous as interim observations are matched by connecting segments where both segments generate identical values /C3/. These three conditions are referred to as the general definition of spline interpolation.

The taxonomy of spline methods follows the type of the $S(t)$ functions.[9] Most often, these functions are polynomials chosen in a way that the derivatives (slopes) of the connecting segments are identical. Generally, a spline's degree and order are determined by the highest exponent of the polynomials in each segment (degree) and by the order of the two derivatives in the connecting segments (order).[10]

Here, we focus on two popular kinds of spline interpolation, in particular. These are:

– linear splines (discussed mostly for didactic reasons), and
– third degree, second order splines.

As for linear splines, let's focus on one interval first, and let this interval be $[t_{k-1}, t_k]$. Let us assume that the values at both ends of the interval are known. If the interpolated curve (stochastic process) between the two end points can be determined, then the missing values can be obtained from this very function.

[8] Here and hereinafter, criteria are abbreviated by "C".

[9] Even though the definition allows the use of different types of functions in each segment, generally the same class of interpolating functions is used in all intervals.

[10] The assumption that a process can be differentiated on a given interval is a serious restriction as it doesn't hold true for Brown motions. Therefore, it is important to emphasize that interpolation techniques are to be applied with caution.

By definition,

$$S_{t_{k-1}}(t_{k-1}) = \alpha_{t_{k-1}} + \beta_{t_{k-1}} t_{k-1} = y_{t_{k-1}},$$
$$S_{t_{k-1}}(t_k) = \alpha_{t_{k-1}} + \beta_{t_{k-1}} t_k = y_{t_k}$$

hold true for any spline. From this system of equations, the unknown parameters ($\alpha_{t_{k-1}}, \beta_{t_{k-1}}$) can be calculated, therefore, the spline can be determined.[11] The solution is identical with the linear estimation discussed earlier, i.e.

$$S^{LIN}(t) = y_t = y_{t_{k-1}} + \frac{y_{t_k} - y_{t_{k-1}}}{t_k - t_{k-1}}(t - t_{k-1}) \quad t \in [t_{k-1}, t_k]. \qquad /3/$$

From this, it follows that the spline parameters can be different in each $[t_{k-1}, t_k]$ interval.

In practice, typically third degree, second order spline interpolations are used, as their flexibility is coupled with a reasonable amount of calculations required.[12] In this case (cubic splines), three more conditions need to be added, extending /C1/–/C3/:[13]

$$S'_{t_{k-1}}(t_k) = S'_{t_k}(t_k), \qquad /C4/$$
$$S''_{t_{k-1}}(t_k) = S''_{t_k}(t_k), \qquad /C5/$$
$$S''(t_1) = 0 \quad S''(t_T) = 0. \qquad /C6/$$

In order to obtain the curves, those represent the interpolation, the parameters in

$$S^{CUB}(t) = S_{t_{k-1}}(t_{k-1}) = y_{t_{k-1}} =$$
$$= \alpha_{t_{k-1}} + \beta_{t_{k-1}}(t - t_{k-1}) + \gamma_{t_{k-1}}(t - t_{k-1})^2 + \delta_{t_{k-1}}(t - t_{k-1})^3 \quad t \in [t_{k-1}, t_k] \qquad /4/$$

[11] The system of equations can always be solved as by definition, $t_k - t_{k-1} > 0$ holds true for the determinant of the coefficient matrix.

[12] Third degree, second order splines are typically referred to as cubic splines.

[13] This representation is valid for natural splines. It is not impossible that the second derivatives at the two endpoints are not equal to zero.

need to be determined, where $4(T-1)$ unknown parameters are paired up with the same number of conditions in /C2/–/C6/, which grants the feasibility of these calculations.[14]

The formerly described cubic spline interpolation is often replaced by CRS[15] in practical scenarios (*Catmull–Rom* [1974]). This method can be applied when non-equidistant time series are presumably equidistant by nature but include missing observations. Let us introduce the notation

$$y_{t_0+\, j\times \Delta} = y_j$$

where y_0 is the first empirical value, y_1 is the second one, etc. The core idea, then, is to assume that all values (observed or missing) fit on a cubic polynomial of known values and derivatives:

$$y(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 \ .$$

This, with respect to the first two points, means

$$\begin{aligned} y(0) &= \alpha_0 \,, \\ y(1) &= \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 \,, \\ y'(0) &= \alpha_1 \,, \\ y'(1) &= \alpha_1 + 2\alpha_2 + 3\alpha_3 \ . \end{aligned}$$

Let us solve the following system of equations for the unknown parameters:

$$\begin{aligned} \alpha_0 &= y(0) \,, \\ \alpha_1 &= y'(0) \,, \\ \alpha_2 &= 3\big[\, y(1) - y(0)\big] - 2y'(0) - y'(1) \,, \\ \alpha_3 &= 2\big[\, y(0) - y(1)\big] + y'(0) + y'(1) \,. \end{aligned}$$

[14] For proof, see *Mészáros* [2011]. The equal number of unknown parameters and conditions is necessary but not sufficient to solve the system of equations, as the non-singularity of the coefficient matrix is also required to obtain an existing and unique solution.

[15] CRS: Catmull-Rom spline.

Plugging the results back into the original polynomial and performing necessary simplifications, we arrive at the following third degree polynomial:

$$y(t) = \left(1 - 3t^2 + 2t^3\right) y(0) + \left(3t^2 - 2t^3\right) y(1) + \\ + \left(t - 2t^2 + t^3\right) y'(0) + \left(-t^2 + t^3\right) y'(1). \qquad /5/$$

The difficulty of solving the equation in /5/ is caused by the fact that the derivative (slope) of the fitted curve is hard to determine. The fundamental concept of the CRS method is the assumption that these derivatives can be obtained based on the observed values. To find the spline on a $\left[\, y_j, y_{j+1} \right]$ interval, let us define the corresponding slopes as

$$y'(j) = \frac{y_{j+1} - y_{j-1}}{2},$$
$$y'(j+1) = \frac{y_{j+2} - y_j}{2}.$$

From this, the third degree polynomial can be rewritten in a matrix form:

$$y(t) = \begin{bmatrix} 1 & t & t^2 & t^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -3 & 3 & -2 & -1 \\ 2 & -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_j \\ y_{j+1} \\ \frac{y_{j+1} - y_{j-1}}{2} \\ \frac{y_{j+2} - y_j}{2} \end{bmatrix}.$$

After basic transformations, we obtain

$$y(t) = \begin{bmatrix} 1 & t & t^2 & t^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -3 & 3 & -2 & -1 \\ 2 & -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} y_{j-1} \\ y_j \\ y_{j+1} \\ y_{j+2} \end{bmatrix},$$

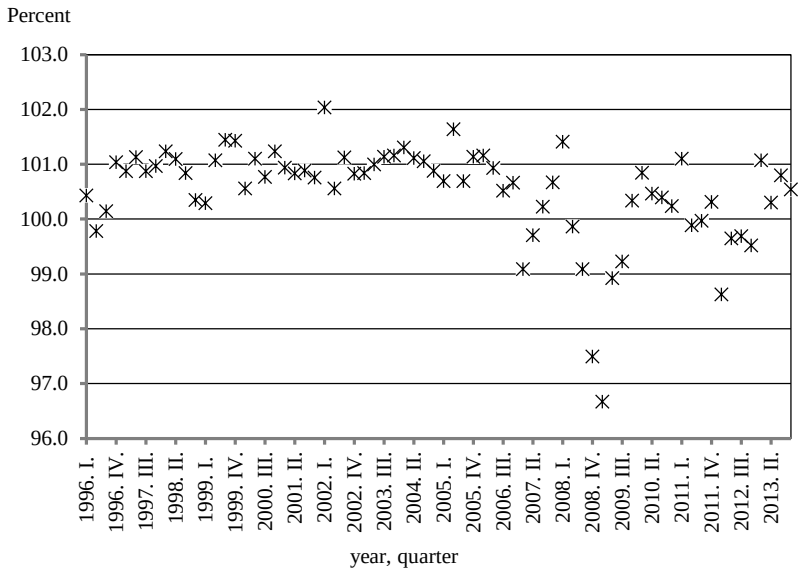
which can also be written as

$$y^{CRS}(t) = \frac{1}{2}\begin{bmatrix} 1 & t & t^2 & t^3 \end{bmatrix}\begin{bmatrix} 0 & 2 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 2 & -5 & 4 & -1 \\ -1 & 3 & -3 & 1 \end{bmatrix}\begin{bmatrix} y_{j-1} \\ y_j \\ y_{j+1} \\ y_{j+2} \end{bmatrix}. \qquad /6/$$

The equation in /6/ is relatively easy to solve, and the results represent the time series between two chosen points. (From the equation, it also follows that the curves resulting from the interpolation can be different in each interval.)

Let us examine this through a simple example. Figure 1 represents the quarterly volume indices of the Hungarian real GDP[16] (quarterly changes) between 1996 and 2013.

Figure 1. Changes in the Hungarian real GDP, 1996–2013
(quarterly volume index)



Source: Hungarian Central Statistical Office (https://www.ksh.hu/docs/eng/xstadat/xstadat_infra/e_qpt008a.html).

In Figure 1, where each quarter is assigned to a value (our time series consists of 72 elements), trends are relatively hard to identify – hence the popularity of line charts.[17] By routinely connecting quarterly observations in such a fashion, we "in-

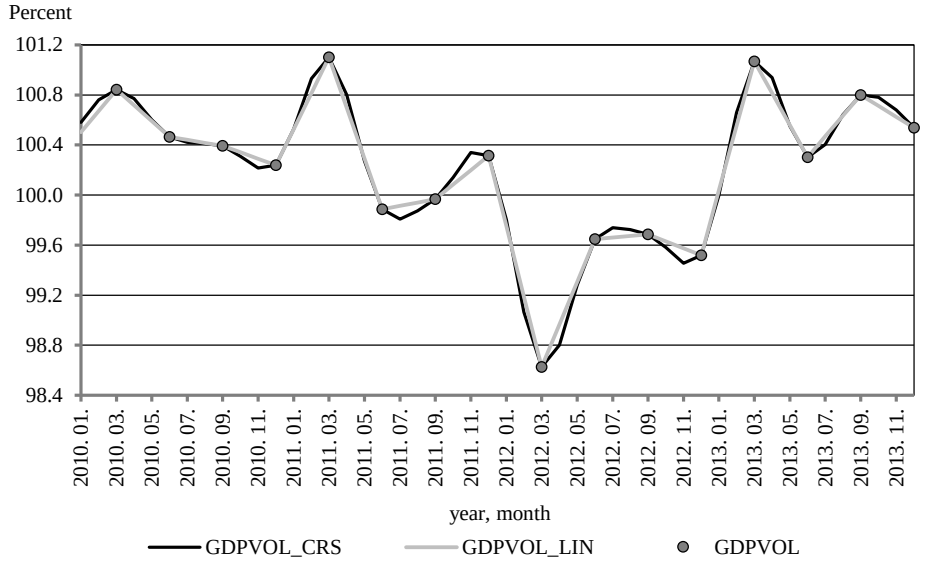
[16] GDP: gross domestic product.

[17] This is one of the reasons why line charts are more often used to represent time series, even if this is somewhat misleading. For rules and principles on visual representation, see *Hunyadi* [2002].

vent" some of the values between actual observations, often without being aware of having done so.

As mentioned earlier, interpolation can be performed in many ways. Figure 2 depicts estimated values by month where data points are broken down from quarterly information, using the linear and the CRS method.[18] In order to give a better visual representation, Figure 2 only includes the last four years' data, so that factual information (GDPVOL) and the values that were obtained from linear and CRS interpolation (GDPVOL_LIN, GDPVOL_CRS) are easier to tell apart.

Figure 2. Changes in the Hungarian real GDP, 2010–2013
(monthly values obtained from interpolated quarterly volume indices)



Source: Here and hereinafter, own calculation.

In certain points, the two interpolated time series in Figure 2 show noticeable differences, and generally, values that were obtained from polynomial splines tend to overshoot the values that were estimated with linear interpolation, especially where trends rebound, such as in mid-2011 or mid-2012. This is one of the reasons why the selection of interpolation techniques requires attention to detail and thoroughness.

[18] The Hungarian Central Statistical Office also provides a monthly breakdown of quarterly GDP values (not volume indices), following a methodologically different path, and therefore arriving at dissimilar numerical results.

2. Falsely identified process characteristics

In order to demonstrate the possible changes in the characteristics of data generating processes that may be induced by interpolation, we turned to simulations.[19] To maintain the comparability of the results, the constants (parameters) were chosen to be identical (or quasi-identical where full identity was not possible). Throughout the analyses, we followed the same principles:

1. For each predetermined model, we generated time series with lengths of 1 000.

2. From these time series, we randomly dropped 10, 20, ..., 90 percent of all "observations".

3. We augmented the resulting non-equidistant time series in two different ways: the missing values were either

– substituted with the expected values of the given process, or

– filled in using cubic interpolation.

4. Finally, based on 1 000 independent "experiments", we examined the differences between the characteristics of the original and the augmented time series for each of the two methods of replacement.

We examined three kinds of data generating processes, each of which are frequently used in empirical time series analysis. From these we generated the following three pairs of time series:

– determined by first order VAR[20] (1) models,
– including stochastic trends (random walk),
– being in perfect first order cointegration.

For all simulated time series, each value preceding the first empirical observation (y_0) was zero, and the random variables were chosen to be white noise processes (denoted by ε_t or by ε_{1t}, ε_{2t} when two processes were being handled concurrently).

The time series that have been generated from the VAR model were based on the model below and following the concept of the Granger test, in order to assess and analyse the unintended changes that may appear in the causal relationships between the time series

$$y_{1t} = 0.9y_{1,t-1} + 0.4y_{2,t-1} + \varepsilon_{1t}\ , \qquad y_{2t} = 0.9y_{2,t-1} - 0.4y_{1,t-1} + \varepsilon_{2t}\ .$$

[19] Simulations were run using EViews 8.1.
[20] VAR: vector-autoregressive.

This, in a matrix form, can be expressed as

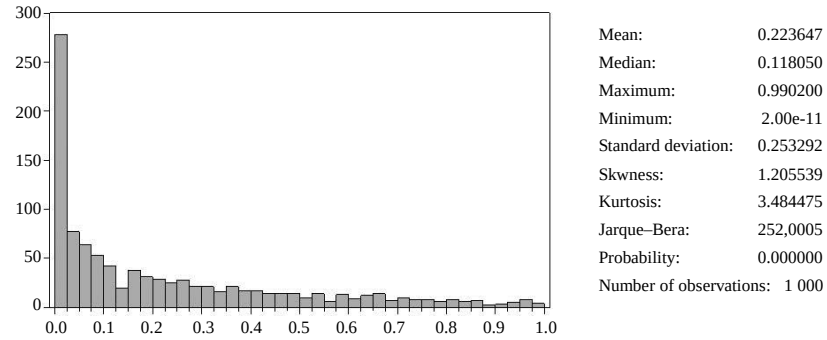
$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 0.9 & 0.4 \\ -0.4 & 0.9 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

Based on common facts, processes that can be formalized using the VAR model are stationary if all eigenvalues of its coefficient matrix are inside the unit circle, and all non-diagonal elements of the parameter matrix are non-zero elements. As both conditions hold true in our example, Granger causality is present between the two variables. In our simulation, we investigated whether it is possible for the causality to disappear following the augmentation of non-equidistant time series.

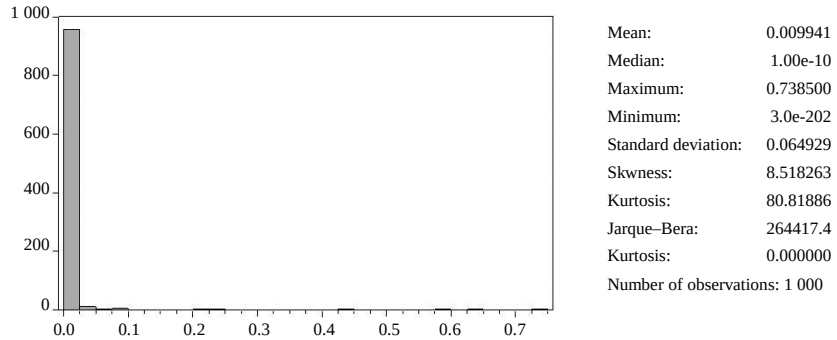
Figures 3. *a*) and 3. *b*) represent the results of the simulations regarding the VAR models.

Figure 3. Levels of significance in Granger tests for causality between the variables obtained from the VAR data generating process

a) If missing values were replaced with their expected values



b) If missing values were replaced using spline interpolation



Note. 90% of the original observations were missing.

Essentially, when missing values were replaced with averages, the Wald test determined causality (rejected the null hypothesis) in only 355 instances at $p = 0.05$. According to the same test at $p = 0.1$, decisions that matched the original data generating process only reached 472. Using spline interpolation, on the other hand, led to a higher number of correct decisions: according to the same test, the number of matching decisions reached 970 and 979 at $p = 0.05$ and $p = 0.1$, respectively, out of 1 000 instances. Based on this, we can conclude that the method of cubic spline interpolation is less likely to identify falsely the data generating process when the number of missing values is substantial.

As for the second type of data generating processes discussed herein, random walk plays a prominent role in time series analysis. Its significance, from our perspective, is due to two particular reasons. Firstly, its presence is the null hypothesis of the unit root tests, and secondly, its shifted version is the pure form of stochastic trends. Accordingly, we simulated two different types of random walk:

– random walk without drift

$$y_t = y_{t-1} + \varepsilon_t ,$$

– random walk with drift

$$y_t = 0.01 + y_{t-1} + \varepsilon_t .$$

Next, we examined whether it is possible to obtain a stationary unit root process if we remove some of its values and then "patch" the gaps using the augmentation techniques described earlier. To test the existence of unit roots, ADF[21] was used.

Regarding the third type of data generating processes discussed herein, we followed the specifications proposed by *Granger* [1988] in our simulations. The processes were:

$$y_{1t} = x_t + \varepsilon_{1t} ,$$
$$y_{2t} = 2x_t + \varepsilon_{2t}$$

where

$$x_t = x_{t-1} + \varepsilon_t .$$

Next, we examined the frequency of occurrences when theoretically cointegrated time series were determined as non-cointegrated, according to the Engle-Granger two-step method.

[21] ADF: augmented Dickey-Fuller test.

The main results of the simulations are summarized in Table 1.

Table 1

Falsely identified data generating processes
(count, out of 1000 simulated instances at $p = 0.05$)

Percent of missing observations	Type of augmentation	Process			
		VAR	RW ($\mu = 0$)	RW ($\mu = 0.01$)	ECM
10	Substitution	0	232	249	0
	Fill-in	0	57	52	0
20	Substitution	0	441	416	1
	Fill-in	0	47	48	0
30	Substitution	0	591	567	8
	Fill-in	0	52	56	0
40	Substitution	0	713	723	33
	Fill-in	0	36	60	0
50	Substitution	0	823	815	38
	Fill-in	0	65	57	0
60	Substitution	0	906	893	34
	Fill-in	0	67	70	0
70	Substitution	9	960	955	44
	Fill-in	0	79	43	0
80	Substitution	151	979	980	20
	Fill-in	1	73	67	0
90	Substitution	645	998	999	9
	Fill-in	30	123	105	1

Note. The term "fill-in", in our table, means that missing values were replaced with their respective expected values. "Substitution" refers to augmentation using cubic spline interpolation. VAR stands for the vector-autoregressive model, RW denotes random walk, and ECM is the acronym denoting the cointegrated system, as it can be operationalized using the error correction mechanism.

The key observations and conclusions, based on the numerical results, are the following:

– Substitution may obscure Granger causality, especially at larger proportions of missing values. Our simulations support the conclusion that spline interpolation leads to better results than using expected values to replace missing observations.

– "Patching" non-equidistant random walk processes with expected values is an unequivocally erroneous choice. The misspecification of the original data generating process using spline interpolation, however, is unlikely, unless the proportion of missing values is substantial. (Note that the augmented Dickey-Fuller test itself results in 50 erroneous decisions out of 1 000 equidistant instances).

– In the case of cointegrated time series, spline interpolation led to fewer (close to zero) misspecifications. According to our analyses, spline interpolation is unquestionably the better alternative.

Based on our simulations, we can state that data augmentation with spline interpolation carries substantially less risk than traditional methods, given that the time series at hand are non-equidistant.

3. Illustrative examples

To understand better the technique of spline interpolation, let us examine two empirical examples.[22]

In our first example, let us take a closer look at the best times of two swimmers, *Daniel Gyurta* (Olympic and world champion, Hungary) and his rival, *Michael Jamieson* (Scotland), particularly their results in the 200-meter breaststroke. Our analysis encompasses the best times in a period of five years, during which 32 competitions were held where either of the swimmers were involved. During this time frame, there were only 7 occasions when both sportsmen participated. For an overview of the raw data set, please refer to Table 2. For a visual representation, see Figure 4.

As we can see, both swimmer's times are "generated" in a non-equidistant, random fashion.[23] Additionally, the dates of relevant competitions do not necessarily coincide, except when both athletes participate in the same meet.

[22] Please note that the results in this section are purely illustrative, and they are not intended to be used for professional decision-making purposes.

[23] The situation would be different if we considered the results of every competition and training, but this falls outside the main purpose of this paper. As our primary goal was to provide an illustrative example of the methods themselves, factors that may influence the swimmers' performance, such as life events etc., have been excluded from our analysis.

Table 2

Best times by meet, 2009–2013
(minutes:seconds.hundreths)

Date	Event	Gyurta	Jamieson
07/26/2009	World Cup	2:08.71	
04/03/2010	British Gas Championships		2:14.85
06/22/2010	British Gas Championships		2:13.63
08/09/2010	European Championships	2:08.95	2:12.73
01/15/2011	Flanders Swimming Cup	2:13.21	2:16.59
10/04/2010	British Commonwealth Games		2:10.97
02/11/2011	BUCS Long Course Championships		2:13.31
03/05/2011	British Gas Championships		2:10.42
03/25/2011	Budapest Open	2:12.67	
06/04/2011	Barcelona Mare Nostrum	2:12.48	2:12.83
06/08/2011	Di Canet Mare Nostrum		2:12.28
06/22/2011	Hungarian Championships	2:10.45	
06/30/2011	Scottish Gas National Open Championships		2:13.04
07/24/2011	World Cup	2:08.41	2:10.54
12/02/2011	Danish Open		2:10.40
01/13/2012	Victorian Age Championships		2:12.15
01/14/2012	Flanders Swimming Cup	2:11.79	
03/03/2012	British Gas Championships		2:09.84
03/29/2012	National Open Championship	2:12.65	
05/21/2012	European Aquatics Championships	2:08.60	2:12.58
06/02/2012	Mare Nostrum		2:11.21
06/13/2012	Budapest Open	2:09.89	
07/06/2012	6th EDF Open Championship		2:11.24
07/28/2012	London Olympics	2:07.28	2:07.43
01/19/2013	Flanders Speedo Cup	2:10.50	
02/08/2013	Derventio eXcel February Festival		2:11.75
03/07/2013	British Gas Swimming Championships		2:10.43
03/29/2013	Budapest Open	2:10.68	
06/13/2013	Sette Colli Trophy	2:10.25	
06/26/2013	Hungarian Championships	2:09.85	
06/28/2013	British Gas Championships		2:07.78
07/28/2013	World Cup	2:07.23	2:09.14

Source: http://bit.ly/1ECuw3N, the official website of the International Swimmers' Alliance. Only best times per meet are considered.

Figure 4. Best times of Gyurta and Jamieson

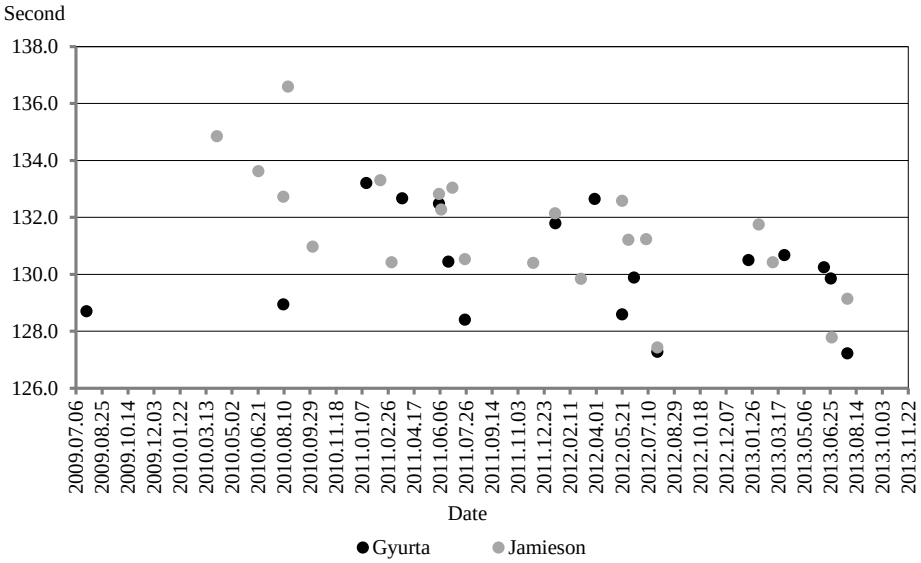


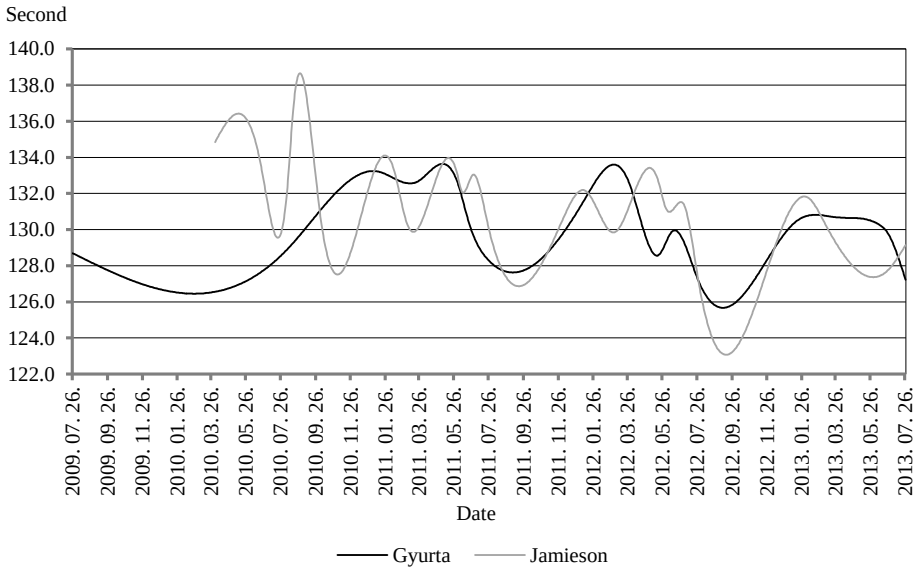
Figure 4 is not overly informative: best times are difficult to identify and comparisons are challenging to make. Even though individual results are obviously easy to compare (e.g. Gyurta beat his rival during the London Olympic Games or at the 2013 World Championship; lesser times are better), but a similar comparison is more difficult to make with respect to the entire time horizon involved in our analysis.

Interpolation, as it provides us with certain values in-between actual observations is one way to make such comparisons possible.[24] Our primary goal, now, is to illustrate the theoretical possibility of estimating changes in performance using interpolation when empirical values are assigned to unequally spaced events. As the dates of competitions were not equally spaced, i.e. the assumption that the time series is equidistant with certain missing observations does not hold true, the application of cubic splines appears to be a reasonable response.[25] For a visual representation of the results, see Figure 5.

[24] Again, it is not our goal to discuss whether the results are actually comparable from a sports professional's point of view (e.g. whether having a competition before, during or after a training camp affects performance, etc.).

[25] This does not imply that competitions are held on a specific day in each month with or without the participation of either Gyurta and/or Jamieson, but it refers to the fact that there are specific sports seasons in each year with longer periods of off-season intervals in-between.

Figure 5. Best times of Gyurta and Jamieson
(estimated values using spline interpolation)



One does not have to be well versed in swimming to realize that the "fictional" values generated by interpolation are not necessarily realistic. Based on this figure, one may come to the conclusion that even though the Scottish athlete has lost the Olympic Finals, his performance has gone through such an improvement during the Olympic Games that he would have been able to achieve significantly better results shortly afterwards, and possibly even set a new world record. Then, following 2012, his performance started to decay again, and despite a rapidly passing improvement, he already got past his top shape before the world championships began. Gyurta, on the other hand, appears to be the athlete with more stable (less unpredictable) numbers whose performance peaks at the most important meets, each year. If we catch ourselves automatically accepting such conclusions as a kind of model verification, we have to remind ourselves repeatedly that spline interpolation is a tool to be handled carefully.

Let us now look at another example that we have referred to earlier in this study. According to the market model (see *Bélyácz* [2009], p. 77.), the yields of a specific stock (or any investment) can be written as

$$r_i = \alpha + \beta r_M$$

where r_i is the yield of the investment *i*, r_M is the yield of the market portfolio, and α and β are model parameters that are to be estimated. In this model, the latter

parameter has a more significant role as it is often used as a proxy, measuring the individual risk of a given investment.

In the specific parameter estimation process, the yield of an individual investment (stock) at time *t* is calculated using

$$r_{it} = \frac{p_{it} - p_{i,t-1}}{p_{i,t-1}} \approx \Delta \log p_{it}$$

where p_{it} denotes the closing price of the given stock on day *t*.

Similarly, market portfolio yields can be estimated using the closing prices of the stock market index (in our example, the BUX[26]). Therefore, we can populate the time series for both variables that are required in our model (market and individual investment yields). Since yields typically form stationary time series, OLS[27] parameter estimation is considered an efficient method.[28]

From our point of view, the fact that daily closing prices are only seemingly equidistant (the values from weekends and holidays are, in fact, missing) is of high importance. In practice, β estimation is often performed after the missing values are replaced with the previous days' closing prices, i.e. yields on such days are made equal to zero. This, however, involves the risk that otherwise existing causal relationships between variables may disappear.

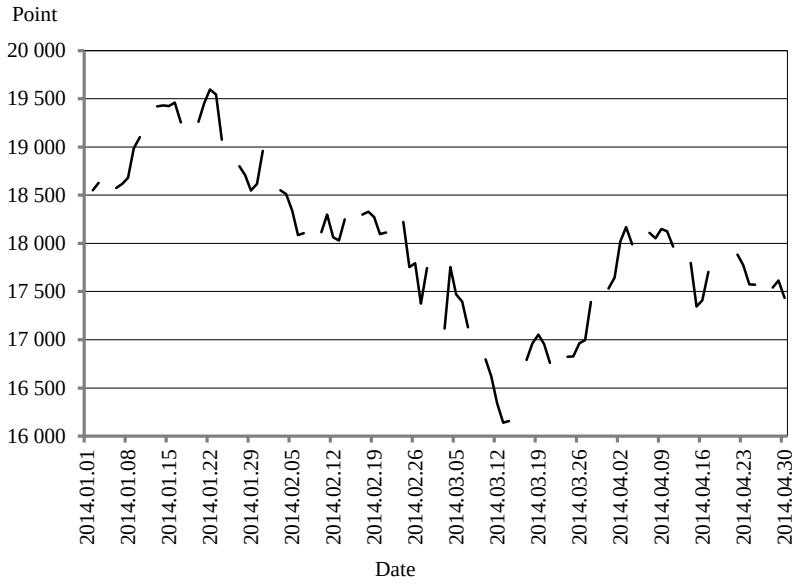
Let us now consider the first four months of 2014, and examine the difference of the results between the two approaches of time series augmentation (fill-in versus substitution). In this example, we are looking at the estimated β values of the blue chips traded on the Budapest Stock Exchange. For a visual representation of the BUX closing prices within the given period, see Figure 6.

[26] BUX: Budapest Stock Index.

[27] OLS: ordinary least squares.

[28] In a previous study (*Varga–Rappai* [2002]), we have shown that parameter estimation leads to more sound results using GARCH (generalized autoregressive conditional heteroscedastic) specifications. This, however, exceeds the focus of this paper.

Figure 6. Closing prices of the BUX in the first four months of 2014



The gaps in Figure 6 represent weekends, holidays and other occasions when trading is suspended. We would obtain similar gaps by depicting the closing prices or daily yields of the aforementioned blue chips, specifically the stock prices of the following companies: MOL, MTELEKOM, OTP, and RICHTER.

We estimated β, the coefficient describing individual investment risk, in three different ways.

– Only considering days of active trading, i.e. with the assumption that yields between Fridays and Mondays are generated the same way as they are between Mondays and Tuesdays. This scenario resulted in 83 individual data points.

– In our second approach, every day within the four-month period was assigned a daily yield, in a way that weekends and holidays generated zero yields, and other values were obtained from actual yields (from the differences between closing prices). This method resulted in a time series consisting of 120 data points for each blue chip.

– Finally, we used cubic spline interpolation in order to eliminate gaps, from which four time series (one for each stock) of 118 data points were obtained, as the first two days of the year had no preceding values.

For a summary of results, see Table 3.

Table 3

Estimated β values by method of augmentation

Blue chip	Approach (length of time series)		
	Weekdays only (T = 83)	0 yield on weekends (T = 120)	Augmented time series using spline interpolation (T = 118)
MOL	0.9109	0.9176	0.9040
MTELEKOM	0.4704	0.4695	0.5017
OTP	1.2715	1.2658	1.6214
RICHTER	1.0825	1.0771	1.2744

The results speak for themselves. On the one hand, replacing missing yields with zeros leads to estimated coefficients that barely differ from estimates, which were obtained by ignoring the existence of weekends and holidays, i.e. by ignoring the effect and role of time at which yields were generated. On the other hand, if we consider the estimated β coefficients that were obtained using spline interpolation, especially the ones that indicate a higher level of risk ($\beta > 1$), we can observe that these estimates are noticeably different from the values that were calculated using either of the first two methods. This implies that changes following weekends and other holidays should be paid special attention. In other words, an additional amount of risk should be associated with individual investments that tend to diverge from the direction of the market as a whole if this divergence is particularly noticeable after weekends and other holidays.

4. Conclusions

One of the most prevalent characteristics of today's information society is the enormous amounts of data that researchers and analysts have to face. Whereas having bigger data sets and longer time series can be beneficial from a certain point of view, one should not forget about the downside of such tremendous amounts of information: the decline in its quality. Data quality is a rather complex category – in this study, the phrase is not used in a way it is defined by statistical terminology, but it is interpreted from the analyst's point of view. From the same perspective, a significant drawback of large data sets and long time series is volatility (statistically speaking, variance or standard deviation). Besides this phenomenon, other difficulties need

to be addressed, such as the appearance of outliers and structural frictions. Many of these phenomena became typical in today's information society, one of which has been addressed herein. One of the main reasons why non-equidistant time series have become more common these days is that analysts now typically rely on multiple data providers, as opposed to the past when a single provider was deemed unquestionable and therefore reliable.

The core dilemma that today's analysts have to face when dealing with such time series is whether they should exclude certain observations in order to qualify them for more traditional modelling techniques by inevitably losing otherwise available information, or if their models should be based on most or all of the available data, taking their unstructured nature, in some way, into account. Since inappropriate choices of data augmentation techniques may affect the basic characteristics of a non-equidistant time series, we have placed significant emphasis on selecting the most desirable methods in our analyses.

The spline interpolation methods discussed herein are relatively straightforward and are supported by the vast majority of standard software packages. Therefore, their use can be recommended to analysts who wish to avoid losing empirical data but also strive to retain the characteristics of the original time series. However, as data augmentation results in conclusions that are not solely based upon original data, time series augmentation should be used with caution. The extent to which analysts should rely on non-empirical data is an ethical question far exceeding the purpose of this study.

References

BÉLYÁCZ, I. [2009]: *Befektetési döntések megalapozása*. Aula Kiadó. Budapest.

BERGSTROM, A. R. [1985]: The Estimation of Parameters in Non-Stationary Higher-Order Continuous-Time Dynamic Models. *Econometric Theory*. Vol. 1. No. 2. pp. 369–385.

BROCKWELL, P. J. [2001]: Continuous-Time ARMA Processes. In: *Shanbhag, D. N. – Rao, C. R.* (eds.): *Handbook of Statistics 19 – Stochastic Processes: Theory and Methods*. Elsevier. Amsterdam. pp. 249–276.

CATMULL, E. – ROM, R. [1974]: A Class of Local Interpolating Splines. In: *Barnhill, R. E. – Reisnfeld, R. F.* (eds.): *Computer Aided Geometric Design*. Academic Press. New York. pp. 317–356.

COCHRANE, J. H. [2012]: *Continuous-Time Linear Models*. NBER Working Paper 5807. National Bureau of Economic Research. Cambridge.

DOOB, J. L. [1953]: *Stochastic Processes*. Wiley. New York.

ENGLE, R. F. [1996]: *The Econometrics of Ultra-High Frequency Data*. NBER Working Paper 5816. National Bureau of Economic Research. Cambridge.

ENGLE, R. F. – GRANGER, C. W. J. [1987]: Co-integration and Error Correction: Representation, Estimation and Testing. *Econometrica*. Vol. 55. No. 2. pp. 251–276.

GRANGER, C. W. J. [1988]: Some Recent Developments in a Concept of Causality. *Journal of Econometrics*. Vol. 39. No. 2. pp. 199–211.

HANSEN, L. P. – SARGENT, T. J. [1991]: Prediction Formulas for Continuous-Time Linear Rational Expectations Models. In: *Hansen, L. P. – Sargent, T. J.* (eds.): *Rational Expectations Econometrics.* Westview Press. Boulder. pp. 209–218.

HUNYADI, L. [1994]: Egységgyökök és tesztjeik. *Szigma*. Vol. 25. No. 3. pp. 135–164.

HUNYADI, L. [2002]: Grafikus ábrázolás a statisztikában. *Statisztikai Szemle*. Vol. 80. No. 1. pp. 22–52.

JONES, R. E. [1985]: Time Series Analysis with Unequally Spaced Data. In: *Hannan, E. J. – Krishnaiah, P. R. – Rao, M. M.* (eds.): *Handbook of Statistics. Vol. 5. Time Series in the Time Domain*. Elsevier. North-Holland. Amsterdam. pp. 157–177.

KOPÁNYI, SZ. [2010]: *A hozamgörbe dinamikus becslése*. PhD-értekezés. Budapesti Corvinus Egyetem. Budapest.

LOMB, R. [1976]: Least-Squares Frequency Analysis of Unequally Spaced Data. *Astrophysics and Space Science*. Vol. 39. No. 2. pp. 447–462.

MÉSZÁROS, J.-NÉ [2011]: *Numerikus módszerek*. GEMAK 6841B. Digitális Tankönyvtár. Miskolci Egyetem. Miskolc.

RAPPAI, G. [2013]: *Bevezető pénzügyi ökonometria*. Pearson. Harlow.

SCHMITZ, A. [2000]: *Erkennung von Nichtlinearitäten und wechselseitigen Abhängigkeiten in Zeitreihen.* WUB-DIS-2011. Wuppertal.

SCHOENBERG, I. J. [1946]: Contributions to the Problem of Approximation of Equidistant Data by Analytic Functions. Parts A, B. *Quarterly Applied Mathematics*. Vol. 4. No. 1. pp. 45–99, pp. 112–141.

VARGA, J. – RAPPAI, G. [2002]: Heteroscedasticity and Efficient Estimates of BETA. *Hungarian Statistical Review*. Vol. 80. Special No. 7. pp. 127–137.

WANG, Z. [2013]: cts: An R Package for Continuous Time Autoregressive Models via Kalman Filter. *Journal of Statistical Software*. Vol. 53. No. 5. pp. 1–19.