Correlation Between Indicators Over Time in Thematic Maps

Matthias Templ

Vienna University of Technology

Abstract: Visualising indicators in thematic maps is nowadays state-of-the-art and many statistical agencies and data providers support their figures also within interactive visualisation. However, mostly raw data values are presented in maps and the visualisation of statistical estimation results is rarely done and topic in this contribution.

For the estimation of cross-correlations of one reference time series to other time series, we show that it is important to prewhiten the time series based on the model estimates of the reference time series. In addition, a simple weighting of time series to increase the importance of recent years over values from the very past is proposed.

Finally, an application of our implemented visualisation tool using European alcohol consumption statistics is shown.

Zusammenfassung: Statistische Ämter und Organisationen unterstützen heute die Präsentation von statistischer Information mit (interaktiven) Grafiken. Jedoch wurden bisher hauptsächlich nackte Zahlen (Totals, . . .) in interaktiven Karten präsentiert, und auf eine Visualisierung von Schätzergebnissen verzichtet, welche Inhalt dieses Beitrages sind.

Die Schätzung von Kreuzkorrelationen von Zeitreihen bezüglich einer Referenzzeitreihe ist erst nach einem *prewhitening* sinnvoll. Zusätzlich wird eine Gewichtung vorgeschlagen, die Zeitreihenwerte in der jüngsten Vergangenheit höher gewichtet.

Anhand von Gesundheitsdaten über den Alkoholkonsum in Europa werden die Schätzungen durchgeführt und mit der klassischen Kovarianzschätzung verglichen. Die Ergebnisse werden entsprechend in Karten dargestellt.

Keywords: Indicators, Prewhitening, Visualisation, Thematic Maps.

1 Introduction

Indicators are collected in large data bases, like the World Bank Database, or data bases from the United Nations, UNESCO, UNIDO, OECD, Eurostat or statistical agencies, providing more and more information each year. Visualisation is therefore an important task and often graphics may help to provide a good overview of the main trends and dependencies of data for both data analysts and policy makers.

Traditionally, raw indicator values are visualised in graphics like line graphs, bar charts, scatterplots and thematic maps. For the latter one, typical interactive representa-

tions are the flash-based cartographic product of Statistics Austria called `i.Map` (Katzlberger, 2011), or the interactive maps of the `OECD eXplorer`[1] for regional statistics.

1.1 Combining Statistical Estimations and Thematic Maps

Gunawardane et al. (2007) introduced new concepts for visualizing indicators: Based on indicators which are available over time for each country, they estimate the correlations between these countries using the Pearson correlation coefficient based on raw data. Interactively, the user then can click on a country on a map. After clicking, the correlations from that country to all other countries are visualised using a color scale for the correlations. Moreover, predictions from simple OLS regression of the values of an indicator of one country to another country and clustering of countries to show the relationships between socio-economic indicators and countries are proposed.

However, the excellent approach to present not only raw indicator values in thematic maps but to show the correlations between countries based on indicator values over time lacks by the estimation of the correlation. The authors propose to use the usual Pearson correlation coefficient (Pearson, 1996) to estimate correlations over time.

1.2 Outline

In the following we show that the estimation of the correlation should be based on state-of-the-art time series methods and we give practical applications using the indicator *alcohol consumption* from the European Community Health Indicators (ECHI) data base[2] for which currently the total amount of alcohol consumption is available from 1970 until 2008 for each country.

In addition, we show that it may also be important to estimate correlations between lagged time series. Finally, we introduce a weighting of the indicator values in order to respect the policy needs, e.g., to consider that actual values are more important than values from the very past.

In the following sections several approaches to estimate correlation are discussed. In Section 2.1 cross-correlation are motivated. However, these estimations should only be applied to stationary time series, which are mentioned in the same section (Section 2.1). To transform the time series adequately, *prewhitening* is used and discussed in Section 2.2. As an extension, weighted correlations are introduced in Section 2.3, where higher weights are given to actual observations. Section 3 illustrates the concepts, and Section 4 concludes.

2 Correlation Between Indicators Over Time

The general approach in this contribution is to estimate correlations of one country to all other countries. In the end, the user has the possibility to click on a country on a map, and then the correlations of this country to all other countries are presented in the map. As

---

[1] `www.oecd.org/gov/regional/statisticsindicators/explorer`
[2] `http://ec.europa.eu/health/alcohol/indicators/index_en.htm`

mentioned before, the Pearson correlation coefficient using the raw data values is not an adequate method to estimate correlations between one time series to all other time series. In the following, we show that measuring correlation after modelling the data provides more meaningful results.

Besides that fact, the time series may be correlated at a specific lag, where the time series is shifted in time. Comparing one time series of a country with a lagged time series of another country, has practical reasons. For example, the economies of Austria and Germany are closely connected to each other, which has historical and cultural reasons. If we assume that the economy in Germany influences the economy in Austria with a delay of one year, a usual correlation may not confirm that the time series are correlated, or produce even misleading results.

To face that problem, the cross-correlation function comes into use. The cross correlation function is a standard method for estimating the degree to which two time series are correlated. However, the cross-correlation should only be estimated from stationary time series.

2.1 Stationarity

Wei (1990) defines multivariate stationary time series as follows:
An m-variate time series $X_t = (x_{t1}, \ldots, x_{tm})^T$ is (weakly) stationary if

- $\mu_X(t)$ is independent of t and
- $\Gamma_X(t + k, t)$ is independent of t for each k,

where $\mu_X = (\mu_{t1}, \ldots, \mu_{tm})^T$ is the vector of means. Γ_X denotes the covariance matrix

$$\Gamma_X(k) = \begin{pmatrix} \gamma_{11}(k) & \ldots & \gamma_{1m}(k) \\ \vdots & \ddots & \vdots \\ \gamma_{m1}(k) & \ldots & \gamma_{mm}(k) \end{pmatrix},$$

consisting of cross covariance functions $\gamma_{ij}(k)$ as defined in the following.

The cross covariance function between x_t and y_t (for $t = 1, \ldots, T$) is defined as

$$\gamma_{xy}(k) = \begin{cases} \mathrm{E}[(x_t - \mu_x)(y_{t+k} - \mu_y)] & \text{for } 0 \leq k \leq T - 1 \\ \gamma_{xy}(-k) & \text{for } -T + 1 \leq k < 0 \end{cases},$$

for $k = 0, \pm 1, \ldots, \pm T$, see Wei (1990) and Brockwell and Davis (1996). Here, μ_x and μ_y are the expectations of the x and y vectors, respectively. k denotes the time difference between the two time series. This leads to the following cross-correlation function:

$$\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y},$$

for $k = 0, \pm 1, \ldots, \pm T$, where σ_x and σ_y are the standard deviations of the vectors x and y, respectively.

If a univariate time series x_t, $t = 1, \ldots, n$, is in focus, the condition "μ_{x_t} is independent of t" is sufficient for stationarity.

2.2   Transformation to White Noise

To compare time series, systematic or deterministic effects over time have to be removed, i.e., first the time series has to be transformed to *white noise*. White noise is a random process, which has mean zero, a constant variance over time, and the covariance equals zero between different time points.

To transform the time series to white noise, different methods can be used. The most common used models for short time series are the ARIMA(1,0,0) and the ARIMA(0,d,0) model (for details about ARIMA models, see, e.g., Box, Jenkins, and Reinsel, 2008).

One possibility is to fit an ARIMA model (autoregressive integrated moving average) to the time series to transform them to (uncorrelated) white noise. There exists a never ending discussion of some researchers about the kind of model to be used for the transformation of the time series to white noise. If indicators are available only for short time series, which is often the case (e.g. yearly estimates from 1970 until 2008), we propose to use only simple models like the ARIMA(0,1,0) and the ARIMA(1,0,0) model. When non-stationary behaviour is expected for the time series, differencing the time series by degree d may induce stationarity, i.e., to apply an ARIMA(0,d,0) model. This is also known as *de-trending*. In the following, Δ denotes the differencing operator, i.e., $\Delta(x_t) = x_{t+1} - x_t$ for $t = 1, \ldots, T - 1$. Hence Δ^d is the differencing operator applied d times to the time series x_t, i.e., $x_{t_d} = \Delta^d X_t$. In practice, d is usually 0, 1 or 2. In general, this approach is common to remove deterministic components of time series.

An ARIMA(1,0,0) model is also often used to tranform the time series to white noise. An ARIMA(1,0,0) equals an autoregression of order 1 (AR(1)). This means that each point of the time series is dependent on the previous. The third option is to apply an ARIMA(1,d,0) model. This equals to apply an AR(1) model to the de-trended data.

Note that Moeller et al. (2003) (mentioned also in Warrenliao, 2005) introduced a distance between two time series for unequally distributed sampling weights,

$$d^2_{STS}(x, y) = \sum_{k=0}^{n_{t-1}} \left(\frac{y_{k+1} - y_k}{t_{k+1} - t_k} - \frac{x_{k+1} - x_k}{t_{k+1} - t_k} \right) .$$

In summary, they propose to use this distance measure – which is in fact the sum of the squared distances between $\Delta(x_t)$ and $\Delta(y_t)$, $t = 1, \ldots, n$, in our case of equal time points – to measure the correlation between short time series instead of Pearson correlations.

To adequately estimate the cross-correlation between a reference time series x_t to other time series, the reference time series has to be made stationary. This is done by applying a certain model to the data to obtain residuals that should now be white noise. However, if a model is separately estimated to all time series, the estimation of the correlation between time series is misleading even if all time series would now be white noise. This is because the dependencies between the time series are corrupted by different models for each time series.

Therefore, Box et al. (2008) proposed to model the first time series and apply the same model to all other time series. Using this concept, the following steps have to be carried out to estimate correlations of a reference time series of one country to time series from the other countries (measuring the same indicator):

1. First an *ARIMA* model is fit to the time series x_t.

2. Transformation of the correlated input series x_t to the uncorrelated white noise series α_t, which consists in fact of the residuals of the fitted time series.

3. The same transformation is applied to all other $j = 2, \ldots, m$ time series using the fitted parameters from point (1), i.e., from modelling x_t, which leads to the processes $\beta_{tj}, j = 2 \ldots, m$.

4. Estimate the cross-correlation between the process α_t and all processes β_{tj}.

2.3   Weighted Correlation Estimation

In some cases the actual values of indicators are of higher interest. Thus a weighted correlation function is needed that assigns higher weights to the recent values and assigns low weights to values in the past.

   Figure 1 illustrates different weighting schemes. The standard correlation is shown by the solid line, where each observation has the same weight. Linearly decreasing weights over time when moving into the past are indicated by the dashed line. The dotted line results in an exponentially decreasing weight.

   Since differentiation is done in many practical cases, let x_{t_d} and y_{t_d}, $t = 1, \ldots, T - d$, define two already differentiated time series. The same method as in the previous section is applied to the time series leading to prewithened time series α_{t_d} and β_{t_d}. Furthermore, the weighting vector $w = (w_1, \ldots, w_{T-d})$ of dimension $T - d$, is known. The sum of the weights should be one ($\sum_{t=1}^{T-d} w_t = 1$). If this is violated, the weights have to be normalized by dividing with the sum of weights. The difference to the previous section is the different handling of the residuals. The focus is on the most recent observations. Lower weights are assigned to residuals from time points in the very past.
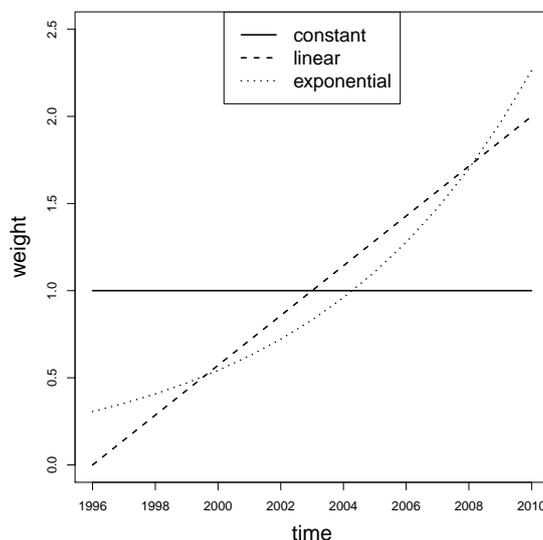

Figure 1: Various weighting functions.

   Figure 1 shows different simple weighting functions for constant weights (classical case), linear increasing weights and exponential increasing weights for a time series.

The next step is the calculation of the weighted mean of the residuals α_{t_d} and β_{t_d}:

$$\bar{\alpha}_{t_d} = \sum_{t=1}^{T-d} w_t \cdot \alpha_{t_d}, \qquad \bar{\beta}_{t_d} = \sum_{t=1}^{T-d} w_t \cdot \beta_{t_d}.$$

After that, the subtraction of those values from residuals of each time series has to be done, i.e., centering them and multiplying the result by the square root of the weights:

$$\alpha_{t_d}^* = \sqrt{w_t} \cdot (\alpha_{t_d} - \bar{\alpha}_{t_d}), \qquad \beta_{t_d}^* = \sqrt{w_t} \cdot (\beta_{t_d} - \bar{\beta}_{t_d}), \qquad t = 1, \dots, T - d.$$

Now the weighted covariance is estimated in the following way:

$$s^2(\alpha_{t_d}, \beta_{t_d}) = \sum_{t=1}^{T-d} \alpha_{t_d}^* \cdot \beta_{t_d}^*.$$

In order to get an unbiased estimation of the covariance, the result has to be divided by the scalar

$$1 - \sum_{t=1}^{T-d} w_t^2.$$

For default weights (all weights have the same value $1/(T - d)$) the conventional unbiased estimate of the covariance with divisor $(T - d)$ is obtained. For estimating the correlations the standard formula is now used:

$$r(\alpha_{t_d}, \beta_{t_d}) = \frac{s^2(\alpha_{t_d}, \beta_{t_d})}{\sqrt{s^2(\alpha_{t_d}, \alpha_{t_d}) \cdot s^2(\beta_{t_d}, \beta_{t_d})}},$$

where $s^2(\alpha_{t_d}, \alpha_{t_d}) = \sum_{t=1}^{T-d} \alpha_{t_d}^{*2}$ and $s^2(\beta_{t_d}, \beta_{t_d}) = \sum_{t=1}^{T-d} \beta_{t_d}^{*2}$.

$r(\alpha_{t_d}, \beta_{t_d})$ is used as the estimation for the correlation between x_t and y_t.

3   Implementation and Examples

The advantages of prewhitening and weighting are first motivated in a simple artificial example where it is known beforehand which time series are correlated. They are presented in Figure 2.

The relation between the first time series (solid line in Figure 2) to all others is of interest in this example as it is typical for the latter correlation plots in maps, where one country is selected and the correlations from any other country to that country are estimated.

An ARMA model is used for prewhitening and it is applied to the first time series (`ts1`). This is realized with the `arma()` function in R, shown in Listing 1.

Listing 1: Fitting an ARMA model to the reference time series.
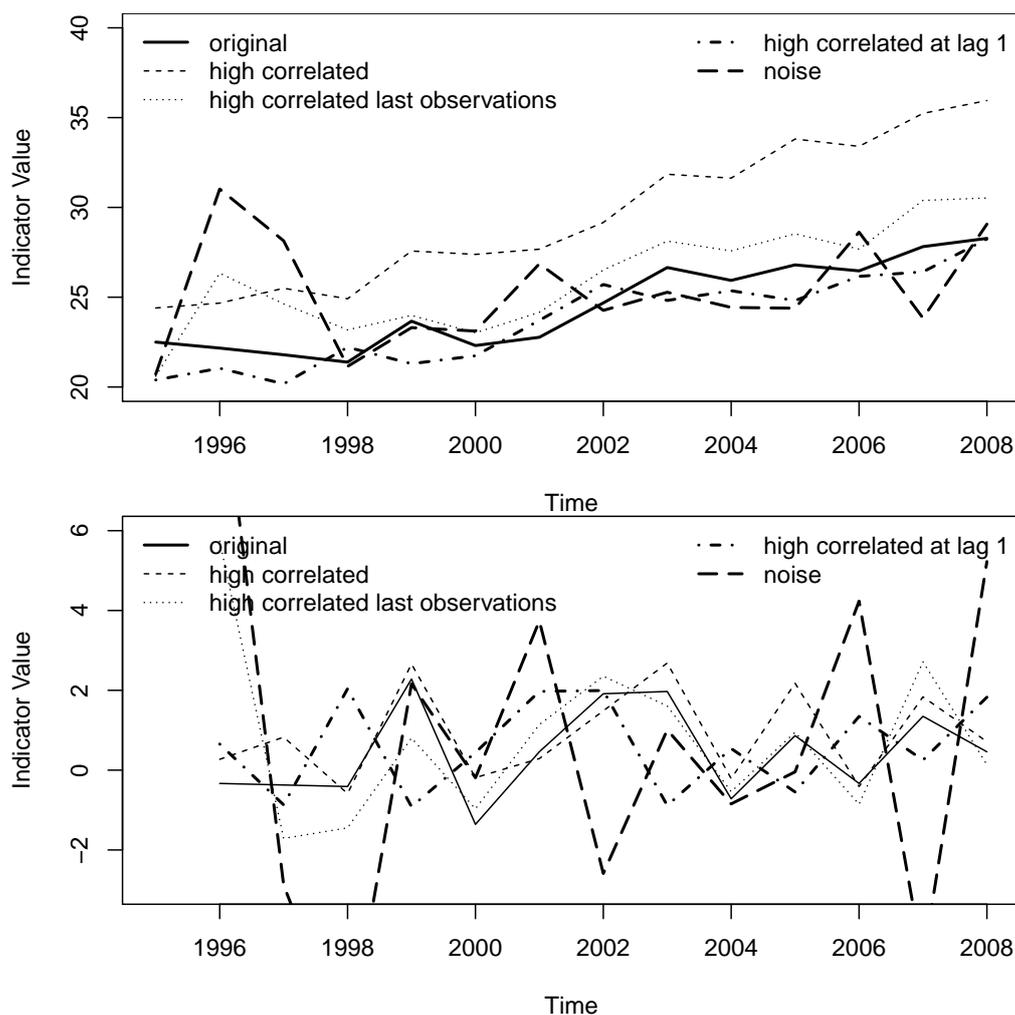
```
ts1a <- arma(ts1)
```

Figure 2: Example for time series, with different correlations to each other. The upper plot presents the original time series, the lower plot the time series after differentiating.

This leads to an ARMA(1,1) model with coefficients $\varphi_1 = 1.068$ for the AR process and $\theta_1 = -0.623$ for the MA process (for details about AR and MA processes, see Brockwell and Davis, 1996). The model parameters are then applied to the other time series by using the `mod.prewhiten()` function (see Listing 2) in the future version of R package **sparkTable** (Kowarik, Meindl, and Templ, 2011), which includes a slightly modified version of the `prewithen()` function in the R package **TSA**.

Listing 2: Prewhitening of the other time series.

```
ts12 <- mod.prewhiten(ts1,ts2,ts1a)
ts13 <- mod.prewhiten(ts1,ts3,ts1a)
ts14 <- mod.prewhiten(ts1,ts4,ts1a)
ts15 <- mod.prewhiten(ts1,ts5,ts1a)
```

As expected, the first time series has a high correlation of 0.883 with the second time series, while the correlation between the first and third time series has a correlation

of 0.40. The correlation between the first and the forth and fifth time series is low as expected, but at lag 1 the correlation is between the first and the fourth time series 0.87.

However, if we assume that the last observations of the time series are more important for policy, the weighting comes into mind. In this example the last observations between the first and the forth time series are higher correlated, which has to be considered. In this example a linearly decreasing weighting function is used, which is then multiplied with the residuals of the two time series after prewhitening, see Listing 3.

Listing 3: Weighted estimation of the correlation between time series.

```
w <- seq(length(ts1))
w <- w/sum(w)
ts14m <- mod.prewhiten(ts1d,ts4d,ts1a)
ccf(ts14m$alpha*w,ts14m$beta*w)
```

This leads to a better result of 0.858 for the correlation between the first and the forth time series at lag 0 (unweighted $= 0.40$). However, this value strongly depends on the weighting function, which should be chosen with care.

3.1 Visualization of the Correlation in Thematic Maps

The function `mapCor()`, which will be included in R-package **sparkTable** (Kowarik et al., 2011), provides a map of Europe and visualises the correlations of a reference country to all other countries. Note that the input for this function is either a matrix or a data frame and the reference country can be selected by just clicking on a country on the map.

The correlations are presented in the map using `hcl`, `rgb` or grey colors. In addition there is an option to use user defined color schemes with the function parameter `"colorScheme"`.

The `mapCor()` function can handle three different representations of the color scale by changing the parameter `chart`, as explained below:

- `"correlation"`: A normal correlation matrix with values between -1 and $+1$. The visualization will use the whole scale.

- `"minmax"`: instead of the whole scale from -1 to $+1$, this option reduces the scale. It reaches from the minimum of the correlation to the maximum of the correlation for the chosen country. This is useful if correlations are close to each other.

- `"distance"`: An option to present a distance matrix, representing a colour scale between 0 and $+1$.

The estimation of the correlation matrix is done by any of the methods mentioned in the previous section. The correlation of an indicator is then shown automatically and the plot updates after picking a single country in the map of Europe.

Interactivity for this function is divided into two parts. On the one hand one can switch between different countries of interest by just clicking on them in the map. The map is then plotted again, using the correlation between the new country to all other countries. On the other hand one can switch between three (respectively four, if a user color scheme was defined) different color schemes and use the one which fits best to the data.

3.2 Application to European Alcohol Consumption Time Series

The proposed methodology is applied to the total consumption of alcohol data from 1970 until 2008, which is one of the most popular indicator from the 88 European Community health indicators provided by the European Commission[3].

Especially, Europe has a problem with alcohol – or without alcohol depending on the views of life. Nevertheless, the EU acknowledges that alcoholic beverages are important economic commodities and they also represent a cultural value for several regions in Europe. Nevertheless, without doubts, the decrease in total alcohol consumption is of high interest for policy makers and for the society. Moreover, it is often of high interest to evaluate the trend of one reference country to the other countries over time. Naturally, the observations from the near past are often of higher interest than the values from the very past for policy makers.

The (graphical) Table 1 shows summary statistics of the total consumption of alcohol of 33 European countries from 1970 until 2008. The *mean*, *min*, *max* and standard deviation (*sdev*) are shown in column 1, 3, 4 and 5. The second column of the table visualizes boxplots based on all values for each country. In the last column, the time series for each country are displayed as sparklines (Tufte, 2006).

The time series originally include 186 missing values (out of 1287 values) which were imputed by regression imputation (response: total alcohol consumption, predictor: time), i.e., a model was fit to each time series and the missing values are predicted by using the fitted model parameters. To not influence the trend, no stochastic error is added to the expected values. The imputations are visible, for example, in the sparkline of Hungary where the values in the past are missing.

Figure 3 is an example for such a plot for which we are interested to look for similar trends of one reference country to all other countries regarding the total alcohol consumption. In left graphic, the Pearson correlations between France and all other countries are presented.

From Table 1 we see that France should have high correlation to Switzerland, Estonia, Italy, Malta and Macedonia, for example. When looking carefully at Table 1 it follows that the time series of Switzerland have higher correlation than Austria that is different especially in the end of the time series. In addition, the total consumption of alcohol over time in Italy is very similar to France, more similar than for Switzerland or Austria.

However, using Pearson correlation coefficients on raw data, these findings are not reflected since Austria, Italy and Switzerland seems to have almost equally high correlation to France, see Figure 3(a).

On the other hand, if a time series model is fit to a reference country and all time series are prewhitened with this model, the results are as expected. This can be seen in Figure 3(b). After prewhitening and weighting of the time series with reference country France, clearly Austria has lower correlation as Switzerland and Switzerland has lower correlation to France than Italy, which is clearly visible in Figure 3(b). As the weighting scheme, simple the linear approach was taken (see Figure 1). Note, that an *arma* or *arima* model may be fit to the reference time series and that the model and model parameters are selected automatically based on the structure of the reference time series.

---

[3]http://ec.europa.eu/health/indicators/echi/list/index_en.htm?echisub=24#echi_3
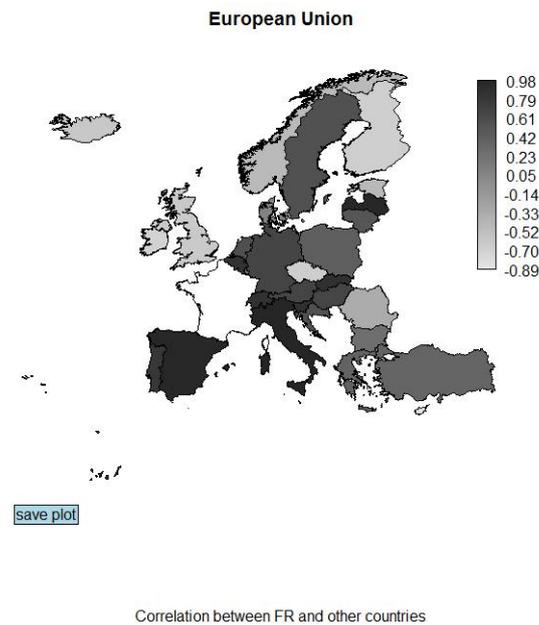
Table 1: Summary statistics and sparklines for the European alcohol consumption data.

	mean	boxplot	*min*	*max*	*sdev*	series
AT	12.81		10.79	16.04	1.35	
BE	11.74		9.69	13.49	1.27	
BG	10.62		7.87	11.79	1.06	
CH	12.52		10.1	14.83	1.38	
CY	7.97		4.1	10.95	2	
CZ	12.81		10.66	15.32	1.41	
DE	13.09		11.62	14.82	0.83	
DK	11.55		8.86	12.78	0.79	
EE	7.01		3.81	16.24	3.22	
ES	14.23		9.67	19.59	3.28	
FI	8.49		5.84	10.45	0.96	
FR	16.46		11.77	21.55	3.17	
GR	9.9		6.74	13.21	1.7	
HR	14.45		11.23	21.6	1.98	
HU	13.19		11.49	14.99	1.05	
IE	11.25		8.57	14.22	1.59	
IS	5.32		4.25	7.53	0.87	
IT	12.53		5.71	19.89	4.25	
LI	9.07		3.62	13.01	2.48	
LT	10.11		5.75	13.63	2.41	
LU	13.89		11.75	15.71	1.04	
MA	10.13		1.44	18.96	5.22	
MT	9.67		1.99	17.72	4.63	
NL	10.38		7.84	12.2	0.89	
NO	5.4		4.55	6.75	0.55	
PL	9.02		7.64	11.49	1.03	
PT	14.93		11.26	20.79	2.35	
RO	10.16		5.1	12.92	1.68	
SE	6.54		4.71	8.06	0.74	
SI	15.42		7.84	23.29	4.65	
SK	11.74		9.27	13.86	1.26	
TR	1.09		-0.62	1.54	0.52	
UK	9.41		6.73	11.78	1.16	

We learnt from this example that the approach of Gunawardane et al. (2007) – to estimate Pearson correlation coefficients on raw time series – fails completely and that prewhitening is essential when dealing with time depended observations to gain reliable results.

European Union

Correlation between FR and other countries
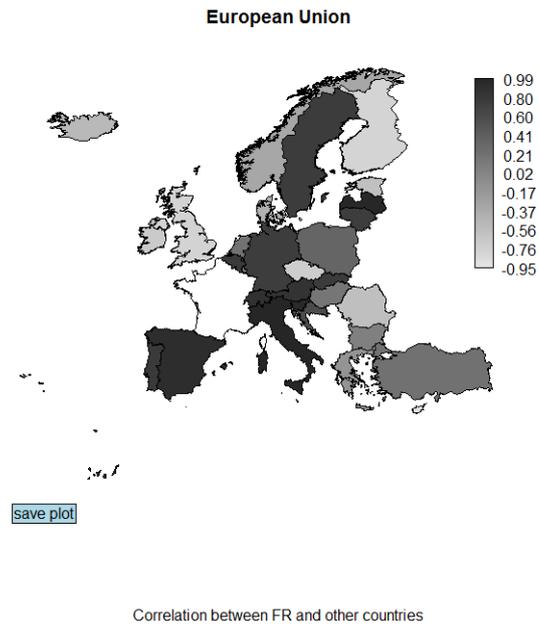
(a) Pearson correlation coefficients based on raw data.

European Union

Correlation between FR and other countries

(b) Correlations coefficients after prewhitening and weighting.

Figure 3: Correlations based on an (interactively) selected reference country (here: France) to all other countries using the total alcohol consumption time series in Europe.

4 Conclusion

In this contribution it is shown on practical examples that time dependent indicator values have to be treated carefully. Applying standard estimators, like the Pearson correlation coefficient as, for example applied in Gunawardane et al. (2007), may lead to wrong conclusions. We show that all time series have to be prewhitened based on the reference time series. In addition to that, we outlined that correlations may also be estimated from lagged time series. However, using estimations from lagged time series includes a detailed analysis of the corresponding time series and may not be applicable when presenting the correlation in maps in an interactive manner.

Weighted estimation of prewithened time series may be prefered whenever the policy makers are more interested in current values than in values from the very past.

The functionality for prewhitening and plotting the indicators in a map is implemented in R and released in future version of the R package **sparkTable** (Kowarik et al., 2011).

Acknowledgement

References

Box, G., Jenkins, G., and Reinsel, G. (2008). *Time series analysis: Forecasting and control*. New Jersey: John Wiley & Sons Inc. (fourth edition)

Brockwell, P., and Davis, R. (1996). *Introduction to Time Series and Forecasting*. New York: Springer.

Gunawardane, P., Feng, J., Lodha, S., Crow, B., Fulfrost, B., and Davis, J. (2007). *Visualizing relationships between global indicators* (Tech. Rep.). UC Santa Cruz: Center for Global, International and Regional Studies.

Katzlberger, G. (2011). *i.Map - interaktive Karten*. Available from `http://www.statistik.at/web\_de/services/interaktive\_karten/index.html` (Version 3.0)

Kowarik, A., Meindl, B., and Templ, M. (2011). *sparkTable: Sparklines and graphical tables for tex and html*. Available from `http://CRAN.R-project.org/package=sparkTable` (R package version 1.2.)

Moeller, C., Klawonn, F., Cho, K.-H., and Wolkenhauer, O. (2003). *Fuzzy clustering of short time series and unevenly distributed sampling points*.

Pearson, K. (1996). Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia. *Philosophical Transactions of the Royal Society A*, *187*, 253–318.

Tufte, E. (2006). *Beautiful evidence*. Cheshire: Graphics Press.

Warrenliao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition*, *38*(11), 1857–1874. Available from `http://dx.doi.org/10.1016/j.patcog.2005.01.025`

Wei, W. (1990). *Time series analysis, Univariate and Multivariate Methods*. Addison-Wesley Publishing Company.

Author's address:

Matthias Templ
Department of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstr. 8-10
1040 Vienna
Austria