Chapter 1

DYNAMIC LINEAR MODELING OF HOMOGENIZED MONTHLY TEMPERATURE IN LISBON

Marco Costa* *and Magda Monteiro*
ESTGA – Escola Superior de Tecnologia e Gestão de
Águeda, Universidade de Aveiro, Portugal
CIDMA – Centro de Investigação e Desenvolvimento em
Matemática e Aplicações

Abstract

This chapter focuses on the statistical modeling of the homogenized monthly average temperature data of Lisbon from 1856 to 2008. An exploratory analysis was performed using linear regression models which indicates the need of considering the temporal dependency and some flexibility in the trend modeling. In order to incorporate the properties of the data it was adopted a dynamic linear models with a fixed effect component. The model was fitted by a two-step procedure which combines the least squares method and the maximum likelihood estimation in the state space framework. The results indicated an average increase of the homogenized monthly temperature in Lisbon in about 0.427°C per century, between 1856 to 2008. Additionally, smoother predictions of the stochastic slopes indicated that the rise of temperature moderately changes according to the month, higher linear increases occurred in the winter months and lower increases occurred in the summer months.

Key Words: Time series analysis, Kalman filter, state space model, temperature data, Lisbon
AMS Subject Classification: 60G35, 62M10, 62M20, 62M05, 93E10.

1. Introduction

This chapter focuses on the long-term time series of monthly temperatures measured in a Portuguese meteorological station of Lisbon. The original dataset contains monthly averages of daily minimum and maximum temperature and their annual means measured at

*E-mail address: marco@ua.pt

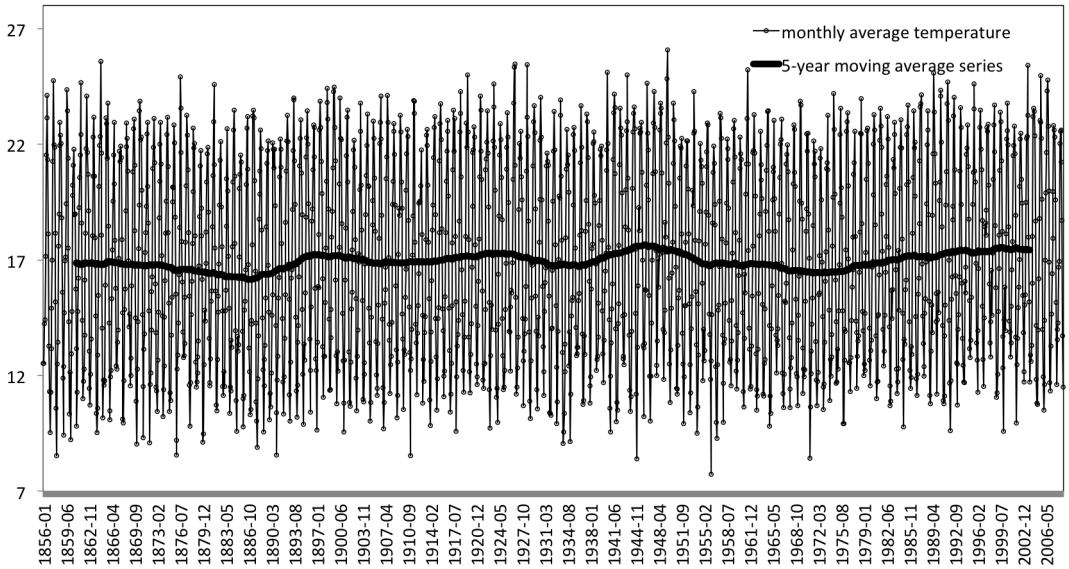

Figure 1. Plot of the homogenized long-term monthly temperature data series in Lisbon from 1856 to 2008. The black line represents the values of the 5-year moving average series.

Instituto Geofísico do Infante D. Luís (IGIDL), from 1856 to 2008. Several studies of climate variability have been done by analyzing global environmental data or based on certain locations. The analysis of Lisbon temperature data was performed by [2] based on a dataset collected at the Climatological Archive of the Portuguese Meteorological Institute and the measurements stretch from January 1856–December 1999. The same dataset were previously considered in the analysis of the evolution of the mean annual temperature in Lisbon by [3]. The results obtained demonstrated a rise in the monthly temperature records in Lisbon, although, with a seasonal pattern. The rise in temperature during the winter months was higher than in summer. Globally, about this dataset was found a mean increase per century in the annual temperature of 1.02428^{o}C.

Sometimes there are several databases on the same region, although they may have been obtained in different locations or resulting from different procedures or methodologies. On the other hand, the quality of data is an important issue which must be discussed mainly when the objective is the environmental changes analysis.

High-quality temperature data sets are vital for climate monitoring, and especially the monitoring of climate change. If a temperature data set is to be used for monitoring climate change it is important that it be homogeneous; that is, changes in the temperature as shown in the data set reflect changes in climate, and not changes in the external (non-climatic) conditions under which the observations are made, [15]. The problem of identification and correction of non-climatic inhomogeneities at the annual and monthly is a well-explored problem (see, [10, 14]).

This chapter focuses on a dataset of homogenized monthly temperature data of Lisbon comprising about 153 years. This dataset was produced by [11] which detected and corrected non-climatic homogeneity breaks in the original dataset. This new data are now avai-

lable for other studies of climate variability, [12]. Relatively to the Lisbon series, two strong non-climatic breaks were detected in the temperature series of Lisbon (IGIDL), which were caused by the changes in the instruments location (1864) and height (1941). These breaks were corrected in the new series.

The data contain monthly extremes and annual means; T_{min} and T_{max} are measured values, temperature range ($DRT = T_{max} - T_{min}$) and T_{aver} are calculated values. In this chapter it is considered the series of the monthly average temperature computed by

$$T_{aver} = \frac{T_{min} + T_{max}}{2}.$$

Figure 1 represents the homogenized long-term monthly temperature data and the 5-year moving average series.

A preliminary analysis will be done in order to detect climate change points that must be considered in the statistical modeling. An analysis will be performed using linear regression modeling which will be considered as a based-model in the time series analysis. The linear regression model will be adjusted considering two deterministic components: trend and seasonality. However, it is expected the existence of a moderate temporal correlation structure, for instance a first order autoregressive process, AR(1), as it is usual in environmental data. To deal with this feature it will be considered dynamic linear models (DLM), which form the class of the so-called state space models and can be viewed as an extension of the usual linear regression models. Besides a comprehensive analysis of the data will also be given attention to the evolution of the temperature in each month, since previous studies have identified that the evolution is not the same for every month of the year. The modeling results will be compared with other works of long-term time series of temperatures.

1.1. Regression linear modeling

A preliminar analysis is performed using linear regression modeling. For simplicity, linear models are largely used as a based-model in time series analysis. Thus, a multiple linear regression model was adjusted considering two deterministic components: a linear trend of the type $\alpha + \beta t$ and a 12-periodic component, $s_t^* = s_{t+12}^*$, corresponding to the seasonal coefficients. This model was fitted considering the formulation

$$Y_t = \beta t + s_t + a_t$$

where $t = 1, 2, ..., 1836$, and such that

$$\alpha = \frac{1}{12}\sum_{i=1}^{12} s_i, \quad s_t^* = s_t - \alpha$$

and a_t is the random error. The random term accommodates the data variability which is not explained by the deterministic components.

The model's parameters are estimated by the least squares method. The results of the overall linear regression model parameters estimation are presented in Table 1. The estimate of the global slope associated to time indicates an average rise in temperature of 0.000283^{o}C per month, or equivalently, in 0.339^{o}C per century. The fit is very good with a

Table 1. Results of the parameters estimation procedure of the overall linear regression model.

Jan	Feb	Mar	Apr	May	Jun
10.72	11.92	13.67	15.36	17.50	20.41
Jul	Aug	Sep	Oct	Nov	Dec
22.31	22.73	21.37	18.26	14.45	11.58
$\widehat{\beta}$	S_a	R^2			
0.000283	1.150	0.9295			

coefficient of determination R^2 of 92.95% and an estimated standard deviation of residuals of 1.150^{o}C. However, a more detailed analysis indicates the existence of a moderate temporal correlation structure, for instance, an autoregressive AR(1) structure, in the residuals as occurred in [2]. Indeed, the series of residuals has a statistical significant serial correlation of 1st order of 0.2739.

If the residuals of a linear regression adjustment are correlated, the estimates of the standard deviations of the coefficients are underestimated and the p-value of the usual t-test are not corrected. However, the temporal correlation can be considered in the usual linear regression models if it is allowed that the errors sequence is, for instance, an autoregressive AR(p) process [1]. In this case, the correlation have impact in the coefficients significance and not in its estimates.

In the work [2] the linear regression models are largely applied to monthly temperature data. As a basis model it is adjusted a regression model with two components both for each month of the year: a linear trend and seasonal coefficients. The linearity of the trend component was suggested in [2] by a graphical analysis of the 12 time series corresponding to each month. This option reflects the conclusion that the restriction of a single constant slope is not suitable for the whole time series. This is also reflected in the time series analyzed in this chapter (see Figure 2).

Indeed, we fitted a simple linear regression model

$$T_{i,t} = \alpha_i + \beta_i t + a_{i,t},$$

where $T_{i,t}$ represents the monthly mean temperature of month i, with $i = 1, 2, ..., 12$, in year $t = i, i+12, ..., 1824+i$ and $a_{i,t}$ represents a sequence of error terms with zero mean, uncorrelated random variables with a constant variance σ_i^2.

Figure 3 presents the intercepts and slopes multiplied by a factor 12×100 obtained in the adjustment of 12 linear regression models, one for each month of the year, to the homogenized long-term monthly temperature data series in Lisbon. The representation of the intercepts estimates shows the temperature seasonal curve in the year, as expected in this type of data. However, the slopes' estimates differ throughout the year. This result is in agreement with [2], although this chapter has been made with another database.

From the results, the largest average monthly rise in the temperature estimated in the data homogenized is 0.841^{o}C per century in January. On the other hand, the estimate of

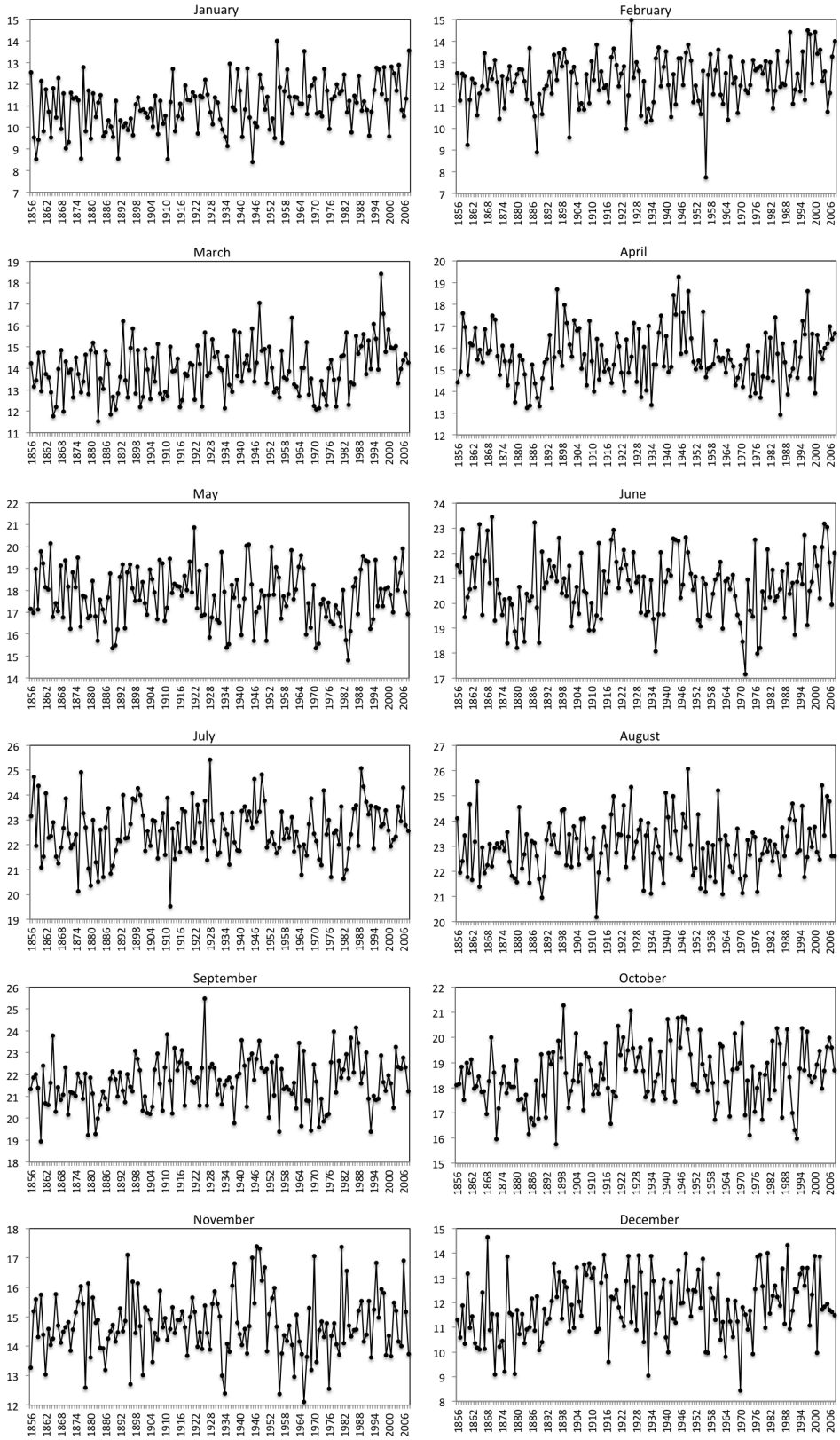

Figure 2. Plot of the 12 time series, one for each month of the year, corresponding to the Lisbon homogenized monthly mean temperature, from 1856 to 2008.

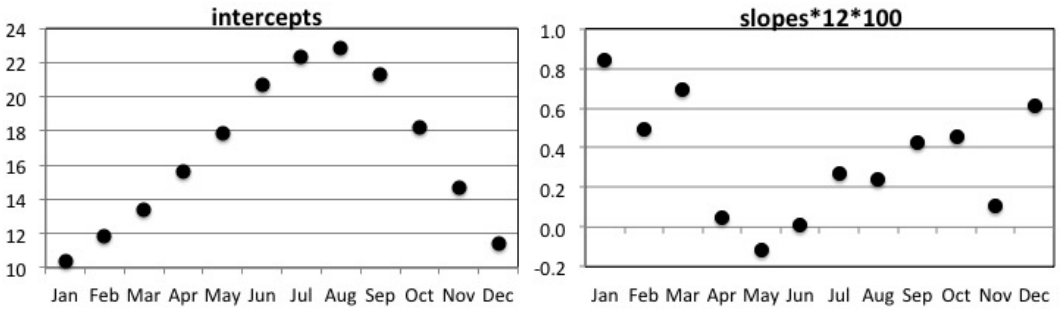

Figure 3. Plot of the intercepts and slopes multiplied by 12×100 to highlight the mean increase per century, resulting from the fit of 12 linear regression models, one for each month of the year, to the homogenized long-term monthly temperature data series in Lisbon from 1856 to 2008.

the slope associated to May is -0.119^{o}C, per century, that is, the months of May have had a decrease, in average, in monthly temperature. Definitely this type of analysis does not take into account the temporal correlation, that although lower between years than between months, is not incorporated in the modeling procedure. For instance, the series of residuals of the linear regressions adjustment show significant first order correlation of 0.117, 0.143, 0.184 and 0.186, respectively to April, May, June and October. Thus, modeling monthly temperatures must allow different degrees of change in the trend component and also temporal correlation usually presented in the monthly climatic data.

In this case, as the main goal is to estimate the trend coefficients, we propose the application of an extension of the linear regression models. Thus, it is adopted a dynamic linear approach to deal with the temporal correlation and the stochastic changes associated to the climatic data without, however, fail to have a simple model to assess possible climate changes.

2. Dynamic linear model

Dynamic linear models are a useful tool to model phenomenons which have temporal dependence and nedd some stochastic dynamics, [9, 5, 6]. In the case of the homogenized monthly temperature, the slope will be considered stochastic allowing the existence of temporal correlation of first order. So, the monthly temperature series is modeled by equations

$$Y_t = t\beta_t + s_t + e_t \quad (1)$$

$$\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t \quad (2)$$

where Y_t is the homogenized monthly temperature variable with $t = 1, 2, ..., 1836$, β_t is the stochastic slope, the *state*, following a 1st order autoregressive process, AR(1), with mean μ. The seasonal behavior is represented by the twelve seasonal coefficients such that $s_t = s_{t+12}$. Random errors e_t and ε_t are uncorrelated white noise sequences such that $E(e_t) = 0$, $E(\varepsilon_t) = 0$, $E(e_t^2) = \sigma_e^2$, $E(\varepsilon_t^2) = \sigma_\varepsilon^2$, for all t, and $E(e_k\varepsilon_r) = 0$, for all k and r.

The model (1)–(2) is very versatile since it can accommodate several statistical properties often presented in environmental data. In fact, the observation equation Eq. 1 can be seen as a regression model and the dependent variable Y_t as a mathematical expectation equal to

$$E(Y_t) = tE(\beta_t) + s_t.$$

When the state process $\{\beta_t\}$ is stationary with mean μ, that is $|\phi| < 1$, it is possible to show some properties. Indeed, if $E(\beta_t) = \mu$, then $E(Y_t) = \mu t + s_t$. Furthermore, if $\{\beta_t\}$ is stationary, then

$$var(\beta_t) = \frac{\sigma_\varepsilon^2}{1-\phi^2} \text{ and } var(Y_t) = t^2 \frac{\sigma_\varepsilon^2}{1-\phi^2} + \sigma_e^2.$$

Since the model has a state space representation it allows obtaining forecast or other predictions of interest (filtered or smoother predictions). In this context, the aim of filtering is to find the expected value of the state vector, β_t, conditional on the information available at time t, that is

$$\widehat{\beta}_{t|t} = E(\beta_t|Y_1, Y_2, ..., Y_t).$$

The aim of smoothing is to take into account the information made available after time t, that is, the mean of the distribution of β_t, conditional on all the sample, denoted by

$$\widehat{\beta}_{t|n} = E(\beta_t|Y_1, ..., Y_n)$$

and is known as a smoother prediction. When the goal is to obtain one-step forecasts to the state or to the observed variable, then the forecast predictions are stated as

$$\widehat{\beta}_{t|t-1} = E(\beta_t|Y_1, ..., Y_{t-1}) \text{ and } \widehat{Y}_{t|t-1} = E(Y_t|Y_1, ..., Y_{t-1}).$$

Since the state process is unobserved, both forecasts and filtered predictions are obtained through the Kalman filter algorithm, while the smoother predictions are computed through the Kalman smoother equations.

Usually, both Kalman filter and smoother are derived based on the assumption that the disturbances e_t and ε_t and the initial state vector $\beta_{1|0}$ are normally distributed. Considering this assumption and all parameters are known, the mean of the conditional distribution of β_t is an optimal predictor of β_t in the sense that it minimises the mean square error (MSE). However, when the normality assumption is dropped, there is no longer any guarantee that both Kalman filter and Kalman smoother will give the conditional mean of the state, [8]. The Kalman filter algorithm and the Kalman smoother predictions, and their MSE, can be found in [13].

The adjustment of the model implies the estimation of parameters

$$\Theta = \{\mu, \phi, \sigma_\varepsilon^2, \sigma_e^2, s_1, s_2, ..., s_{12}\}.$$

Usually, in the context of the Gaussian state space models, parameters are estimated through the EM-algorithm combining two steps: the expectation (E) step and the maximization (M) step, [13]. The EM algorithm is an iterative method for finding maximum

likelihood estimates. However, the iterative equations of this algorithm must be modified to the specific model (1)–(2). When normality is not guaranteed or is not appropriate, other estimation approaches can be adopted as the distribution-free estimators ([4, 7]).

Since the model considered in this chapter is the usual state space model added with a fixed component, the seasonal coefficients, the parameters estimation is performed through a two-step procedure which applies the least squares method for the fixed effects, $\Theta_1 = \{s_1, s_2, ..., s_{12}\}$, and the EM algorithm to estimate the parameters associated to the usual state state model $\Theta_2 = \{\mu, \phi, \sigma_e^2, \sigma_\varepsilon^2\}$.

So, the two-step procedure consists on the following algorithm:

1. Initiate the fixed effects parameters, $\widehat{\Theta}_1^{(1)}$, with the least squares estimates from the based-model (presented in Table 1);
2. Compute residuals $\widehat{\eta}_{t,1}^{(1)} = Y_t - \widehat{s}_t^{(1)}$;
3. Consider $\widehat{\eta}_{t,1}^{(1)} = Y_t - \widehat{s}_t^{(1)}$ and obtain the maximum likelihood estimates of Θ_2, $\widehat{\Theta}_2^{(1)}$, through the observation equation

$$\widehat{\eta}_{t,1}^{(1)} \quad = \quad t\beta_t + e_t$$

4. Reconstruct $\widehat{\Theta}^{(1)} = \widehat{\Theta}_1^{(1)} \cup \widehat{\Theta}_2^{(1)}$;
5. Compute the residuals $\widehat{\eta}_{t,2}^{(1)} = Y_t - t\widehat{\beta}_{t|t-1}^{(1)}$;
6. Consider $\widehat{\Theta}^{(i)} = \widehat{\Theta}_1^{(i)} \cup \widehat{\Theta}_2^{(i)}$ the estimates in the iteration i.
   (a) Compute $\widehat{\Theta}_1^{(i+1)}$ by the least square method considering the regression model $\widehat{\eta}_{t,2}^{(i)} = s_t + a_t$;
   (b) Compute the residuals $\widehat{\eta}_{t,1}^{(i+1)} = Y_t - \widehat{s}_t^{(i+1)}$;
   (c) Consider $\widehat{\eta}_{t,1}^{(i+1)} = Y_t - \widehat{s}_t^{(i+1)}$ and obtain the maximum likelihood estimates of Θ_2, $\widehat{\Theta}_2^{(i+1)}$, through the observation equation

$$\widehat{\eta}_{t,1}^{(i+1)} \quad = \quad t\beta_t + e_t$$

   (d) Reconstruct $\widehat{\Theta}^{(i+1)} = \widehat{\Theta}_1^{(i+1)} \cup \widehat{\Theta}_2^{(i+1)}$;
   (e) Compute the residuals $\widehat{\eta}_{t,2}^{(i+1)} = Y_t - t\widehat{\beta}_{t|t-1}^{(i+1)}$;
   (f) If $\widehat{\Theta}^{(i+1)}$ verifies a convergence condition, for instance $||\widehat{\Theta}^{(i+1)} - \widehat{\Theta}^{(i)}|| < \delta$, then stop the iterative process, else return to 6. a).

When the convergence condition is hold the parameter vector Θ is estimated by $\widehat{\Theta}^{(i+1)}$. Approximated standard deviations can be compute in each step by the usual procedures.

Table 2. Estimates and approximated 95% confidence intervals of the parameters of the dynamic linear model.

parameter	estimate	lower bound	upper bound
μ	3.555E-04	3.499E-04	3.611E-04
ϕ	0.978	0.977	0.978
σ^2_ε	1.162E-08	1.145E-08	1.179E-08
σ^2_e	1.142	1.140	1.144
Jan	10.636	10.470	10.801
Feb	11.838	11.672	12.004
Mar	13.597	13.431	13.762
Apr	15.283	15.118	15.449
May	17.426	17.260	17.592
Jun	20.340	20.174	20.506
Jul	22.242	22.076	22.407
Aug	22.658	22.492	22.824
Sep	21.291	21.126	21.457
Oct	18.182	18.016	18.348
Nov	14.379	14.213	14.545
Dec	11.500	11.334	11.666
$\ln Log$	-2796.41		

3. Results

Table 2 presents the estimates and the approximated 95% confidence intervals of the parameters of the model (1)–(2) and Figure 4 represents both the homogenized long-term monthly temperature data series in Lisbon and smooth predictions from 1856 to 2008.

The autoregressive estimate verifies the stationary condition, $|\widehat{\phi}| < 1$, of the state process $\{\beta_t\}$, although close to 1. This model has a coefficient of determination of $corr^2(Y_t, \widehat{Y}_{t|t-1}) = 93.40\%$.

The model verifies the assumption of normality which was checked by the residuals distribution, $\widehat{\eta}_t = Y_t - \widehat{Y}_{t|t-1}$. Indeed, Figure 5 shows the histogram and the Q-Q normal plot of residuals $\widehat{\eta}_t$. Additionally, the normal distribution is not rejected for both Kolmogorov-Smirnov and Shapiro-Wilk tests (with p-values equal to 0.200 and 0.458, respectively). The proposed model incorporates a significant part of the temporal correlation identify in the overall regression model. Indeed, the first order correlation of residuals decreases from 0.2739 to 0.1574.

As the state process $\{\beta_t\}$ represents a stochastic slope of a global linear trend, the estimate of μ is the global average of the linear trend and it has a special interpretation in this context. Attempting that data are monthly observed, the estimate of μ indicates an average rise of the homogenized monthly temperature in Lisbon in about 0.0003555°C per month, or equivalently, in 0.427°C per century, between 1856 to 2008. With 95% of confidence, this value is within the confidence interval of $[0.41987, 0.43328]$, that is, the

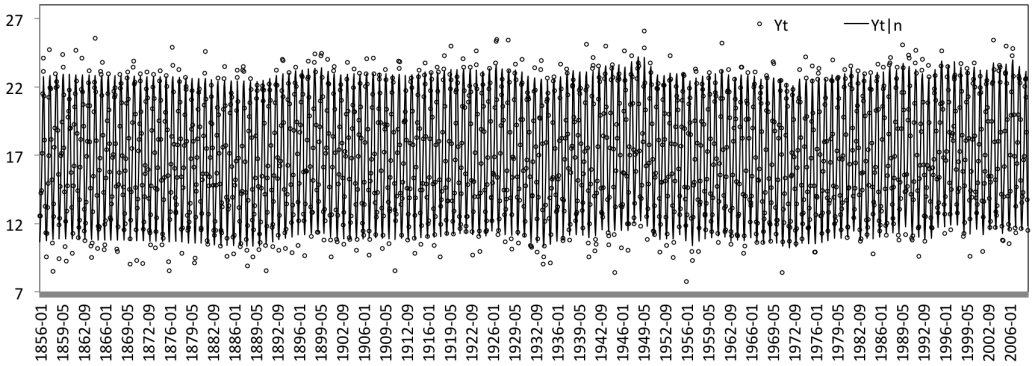

Figure 4. Plot of the both the homogenized long-term monthly temperature data series in Lisbon and smooth predictions from 1856 to 2008.

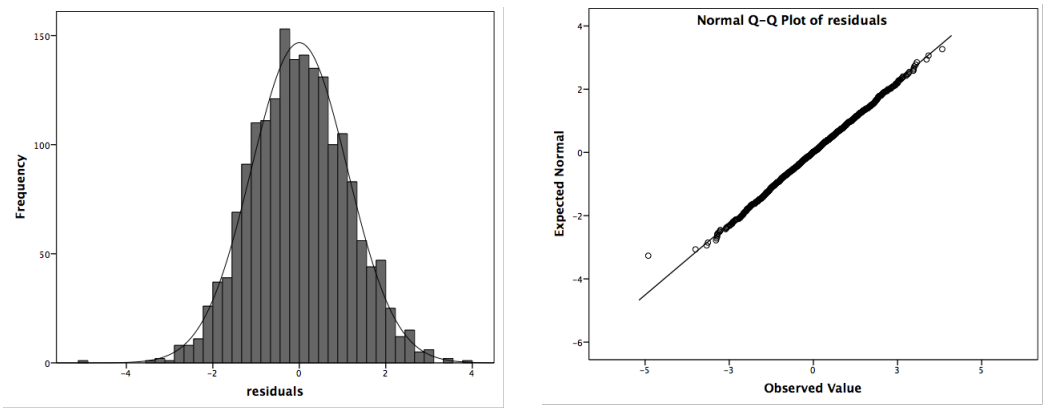

Figure 5. Histogram (left) and normal Q-Q plot of residuals from the adjustment of the dynamic linear model.

slope process has a statistically significant average.

Notice that the initial overall regression model had estimate the century rise temperature in about 0.339°C, that is, the adjustment of the dynamic linear model updates the global rise of the temperature in 26%.

The analysis of Kalman smoother predictions of the state process, $\widehat{\beta}_{t|n}$, is relevant to investigate the behavior of the trend's slope in a more small time scale. Indeed, these predictions indicates an accurate prediction of the local linear trend in each month take into account all sample. Figure 6 represents the smoother predictions of the stochastic slope, $\widehat{\beta}_{t|n}$. This plot suggests that there is a structural break around the year of 1889. So an analysis of these predictions must be taking into account this fact and future research must studies this phenomenon.

Since the proposed model allows different slopes for each month, the global averages of the $\widehat{\beta}_{t|n}$ for each month of the year were computed to investigate a possible pattern in the rise of the homogenized temperature over the year. So, Figure 7 shows the monthly averages of the smoother predictions of slopes for each month of the year converted into a

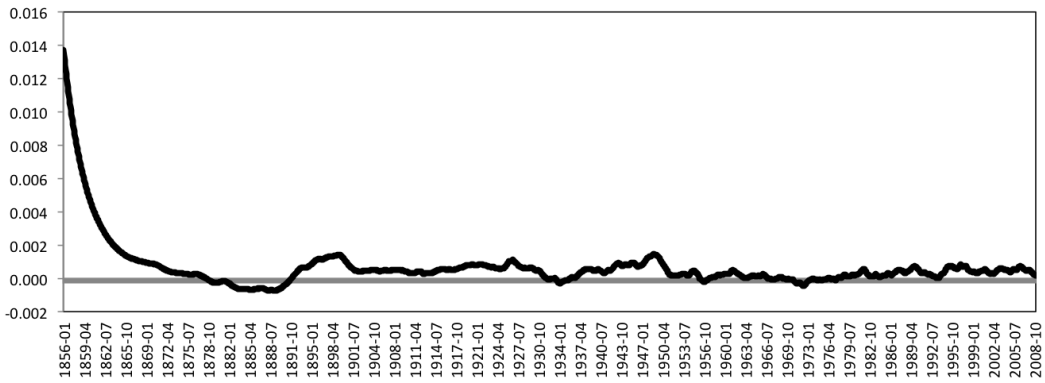

Figure 6. Plot of the smoother predictions of the stochastic slope β_t, $\widehat{\beta}_{t|n}$.

century, in the last century (1909 to 2008).

Results show that in the last century the homogenized temperature in Lisbon has risen in all months of the year. However, there is an annual pattern: greater rises were estimated in the winter months whereas smaller rises were verified in the summer months. The major average rise of temperature is in January and the lower is in June. So, in the last century the winter months were being less cold with a greater monthly rate than the summer months were being more hot.

4. Conclusion

The analysis performed in this chapter shows that dynamic linear models are suitable to model temperature series and they allow to obtain pertinent findings in term of climate change point of view. On the one hand, the main advantage of DLM is its structure which allows incorporating both temporal correlation presented in environmental data and an intrinsic stochastic behavior of this type of data. One the other hand, the flexibility of DLM enables also to consider fixed effects which usually are treated by the regression linear modeling approach. At the same time, the DLM has associated both the Kalman filter and the Kalman smoother to obtain accurate predictions and forecasts.

Data analyzed in this chapter are a long-term series of monthly temperature in Lisbon in a considerable time window. This allowed to fit the proposed model and its statistical validation.

Globally, this model estimates the average increase of the homogenized monthly temperature in Lisbon in about 0.427^{o}C per century, between 1856 to 2008. However, temperature rise was not the same in all the months of the year. Indeed, the highest growth rates were obtained in the winter months while the lower growth rates, although quite significants, were estimated in the summer months.

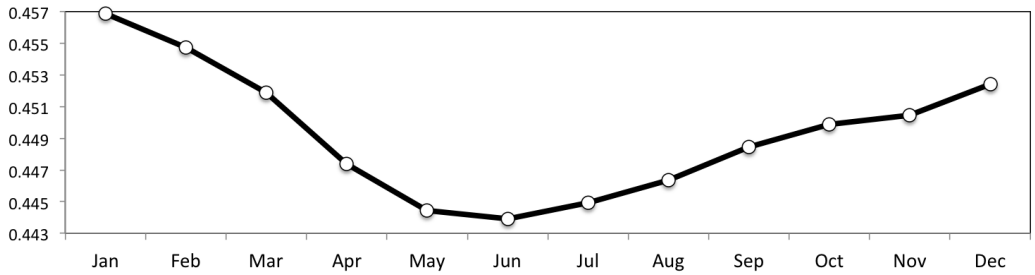

Figure 7. Plot of the century monthly average rise of the homogenized temperature in the last century of the sample.

Acknowledgements

Authors were partially supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology ("FCT– Fundação para a Ciência e a Tecnologia"), within project UID/MAT/04106/2013.

References

[1] T. Alpuim, A. El-Shaarawi, On the efficiency of regression analysis with AR(p) errors, *Journal of Applied Statistics* **35**: 717–737 (2008).

[2] T. Alpuim, A. El-Shaarawi, Modeling monthly temperature data in Lisbon and Prague, *Environmetrics* **20**: 835–852 (2009).

[3] S. Antunes, *Caracterização da Variabilidade Climática Interanual em Portugal Continental*, M.Sc. Thesis University of Lisbon, Lisbon, 1998.

[4] M. Costa, T. Alpuim, Parameter estimation of state space models for univariate observations, *Journal of Statistical Planning and Inference*, **140**: 1889–1902 (2010).

[5] M. Costa, M. Monteiro, A.M. Gonçalves, Kalman Filtering Approach in the Calibration of Radar Rainfall Data: A Comparative Analysis of State Space Representations, In *Rainfall: Behavior, Forecasting and Distribution*, ed. Olga E. Martn and Tricia M. Roberts, ISBN: 978-1-62081-591-5, Nova Science Publishers, New York, 2012.

[6] M. Costa, M. Monteiro, A mixed-effect state space model to environmental data, *Proceedings of the ICNAAM-2014*, *AIP Conference Proceedings*, **1648**: 110002-1–110002-4 (2015).

[7] A.M. Gonçalves, M. Costa, Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering, *Stochastic Environmental Research and Risk Assessment* , **27**: 1021–1038 (2010).

[8] A.C. Harvey, *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, 1996.

[9] P. Kokic, S. Crimp, M. Howden, Forecasting climate variables using a mixed-effect state-space model, *Environmetrics* **22**: 409–419 (2011).

[10] S. Li, R. Lund, Multiple changepoint detection via genetic algorithms, *Journal of Climate* **25**: 674 – 686 (2012).

[11] A.L. Morozova, M.A. Valente, Homogenization of Portuguese long-term temperature data series: Lisbon, Coimbra and Porto, *Earth Syst. Sci. Data* **4**: 187–213 (2012).

[12] A.L. Morozova, M.A. Valente, Homogenization of Portuguese long-term temperature data series: Lisbon, Coimbra and Porto, doi:10.1594/PANGAEA.785377 (2012).

[13] R.H. Shumway, D.F. *Time series analysis and its applications : with R examples*, Springer, New York, 2006.

[14] A. Toreti, F.G. Kuglitsch, E. Xoplaki, J. Luterbacher, H. Wanner, A novel method for the homogenization of daily temperature series and its relevance for climate change analysis, *Journal of Climate* **23**: 5325–5331 (2010).

[15] B. Trewin, A daily homogenized temperature data set for Australia, *International Journal of Climatology* **33**: 1510–1529 (2013).