engineering proceedings

MDPI

Proceeding Paper

Improving Predictive Accuracy in the Context of Dynamic Modelling of Non-Stationary Time Series with Outliers †

Fernanda Catarina Pereira [1,*], Arminda Manuela Gonçalves [2,‡] and Marco Costa [3,‡]

1 Centre of Mathematics, University of Minho, 4710-057 Braga, Portugal
2 Department of Mathematics and Centre of Mathematics, University of Minho, 4710-057 Braga, Portugal
3 Centre for Research and Development in Mathematics and Applications, Águeda School of Technology and Management, University of Aveiro, 3810-193 Aveiro, Portugal
* Correspondence: id9976@alunos.uminho.pt

† Presented at the 9th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 12–14 July 2023.

‡ These authors contributed equally to this work.

Abstract: Most real time series exhibit certain characteristics that make the choice of model and its specification difficult. The objective of this study is to address the problem of parameter estimation and the accuracy of forecasts *k*-steps ahead in non-stationary time series with outliers in the context of state-space models. In this paper, three methods for detecting and treating outliers are proposed. We also present a comparative study of the proposed methods using data simulated from a local level model with sample sizes of 50 and 500 and with various combinations of parameters, with a 5% contamination error rate of the observation equation. The results were evaluated in terms of the accuracy of model parameters and the forecasts *k*-steps ahead, as well as the detection rate of true outliers. These methodologies are applied to three real examples. This study shows that the local level model is sufficiently robust even for non-stationary contaminated series, in the sense that they are able to handle non-stationary time series and outliers in a satisfactory way.

Keywords: outliers; contaminated data; non-stationary time series; state-space models; Kalman filter; simulation study

Citation: Pereira, F.C.; Gonçalves, A.M.; Costa, M. Improving Predictive Accuracy in the Context of Dynamic Modelling of Non-Stationary Time Series with Outliers. *Eng. Proc.* **2023**, *39*, 36. https://doi.org/10.3390/engproc2023039036

Academic Editors: Ignacio Rojas, Hector Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela

Published: 29 June 2023

1. Introduction

State-space models were originally developed in aerospace engineering in the early 1960s for the purpose of monitoring and correcting the trajectory of a spacecraft headed to the moon. Today, these models have wide applicability in many areas, such as finances [1], ecology [2], machine learning [3], and time series analysis and forecasting [4–7]. These models, associated with the Kalman filter algorithm [8], are a very powerful tool given their ability to update predictions both in real time and in a recursive procedure as new observations of the time series become available, thus improving the accuracy of predictions. In addition, state-space models are very flexible due to their ability to incorporate fixed effects and stochastic components that can represent the different unobserved components, such as periodic structures, trends, seasonality, and temporal correlation. These components describe the structural variation of the time series under study. Furthermore, potential covariates can be added because they are important to explain the process and complement the information introduced by the different stochastic components of the model. These models include two sources of variability: one corresponding to measurement errors and the other to process variations. In this way, it becomes simpler to interpret both errors separately. One advantage of these models is that they do not require the assumption of stationarity and can handle time series with missing values in a particularly simple way [4,9]. However, the existence of outliers in real data can influence the estimation and prediction accuracy of both the parameters.

Outliers can be a problem for model specification and prediction accuracy, since the Kalman filter is not generally robust to the presence of outliers. An incorrectly specified model can lead to incorrect covariance matrices of predictions given by the Kalman filter, and thus there is no way to describe the actual quality of the filter [10]. According to [11], the presence of outliers in a time series can induce non-Gaussian heavy-tailed noise, leading to misspecified models, biased estimates, and inaccurate forecasts. The authors of [12] showed that simple linear Gaussian state-space models can present estimation problems. Therefore, in this paper, several methods of detecting and treating outliers are discussed. These methods will be compared and illustrated with a simulation study that considers a simple Gaussian stationary state-space model with 5% data contamination. To create the non-stationarity scenario, the local level model, which is a particular case of the state-space model, will be considered for the sake of simplicity. Detection and treatment of the methods' performance is evaluated by the root-mean-square error (RMSE) and the mean absolute error (MAE) of the Gaussian likelihood of the parameters' estimates and the one-step ahead predictions of the time-series variable. Several scenarios are considered accounting for different combinations of parameters and times series sizes, n in this specific case, (n = 50,500). Time series simulations are generated until 1000 time series have a state-space model with valid estimates, i.e., estimates within the space parameter.

2. Methodologies

The univariate state-space model can be represented by the observation and state equations, respectively, given by

$$Y_t = W_t \beta_t + e_t \tag{1}$$
$$\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t \tag{2}$$

where t represents the time, Y_t is the observed data, W_t is a factor assumed to be known that relates the observation Y_t to the latent variable β_t at time t. The disturbances e_t and ε_t are independent and identically distributed, with Gaussian distribution of zero mean and variances σ_e^2 and σ_ε^2, respectively, and are uncorrelated with each other.

The state β_t is a latent variable and therefore must be estimated. The Kalman filter algorithm ([8]) provides optimal unbiased linear one-step ahead and update estimators of the unobservable state β_t. Let $\Theta = \{\phi, \sigma_e^2, \sigma_\varepsilon^2\}$ be the vector of the model's unknown parameters, let $\hat{\beta}_{t|t-1}$ denote the predictor of β_t based on the observations $Y_1, Y_2, \ldots, Y_{t-1}$ and $P_{t|t-1}$ be its mean square error, i.e., $E[(\hat{\beta}_{t|t-1} - \beta_t)^2]$. The one-step ahead forecast for the observable vector Y_t is given by $\hat{Y}_{t|t-1} = W_t \hat{\beta}_{t|t-1}$. When, at time t, Y_t is available, the prediction error or innovation, $\eta_t = Y_t - \hat{Y}_{t|t-1}$, is used to update the estimate of β_t (filtering) through the equation

$$\hat{\beta}_{t|t} = \hat{\beta}_{t|t-1} + K_t \eta_t,$$

where K_t is called the Kalman gain and is given by $K_t = P_{t|t-1} W_t (W_t^2 P_{t|t-1} + \sigma_e^2)^{-1}$. The mean square error of the updated estimator $\hat{\beta}_{t|t}$, represented by $P_{t|t}$, verifies the relationship $P_{t|t} = P_{t|t-1} - K_t W_t P_{t|t-1}$. Furthermore, the predictor of β_{t+k} at time t is given by

$$\hat{\beta}_{t+k|t} = \mu + \phi^k \left(\hat{\beta}_{t|t} - \mu \right),$$

and its mean square error is $P_{t+k|t} = \phi^{2k} P_{t|t} + \sum_{i=0}^{k-1} \phi^{2i} \sigma_\varepsilon^2$.

Outlier Detection and Treatment Procedures

Three approaches to outlier detection and treatment are presented. The first approach is based on linear interpolation, which represents the naive method. The other two approaches are based on iterative processes from the robust Kalman filter and from the Kalman filter in the missing values perspective.

1. Linear interpolation (LI)
    - Outlier detection: Observations are considered outliers if they are less than $Q_1 - 1.5IQR$ or greater than $Q_3 + 1.5IQR$, where Q_1 and Q_3 denote the first and third quartiles, respectively, and IQR (interquartile range) is the difference between the third and first quartiles (IQR rule).
    - Outlier treatment: Any outliers that are identified are replaced by LI using the neighbouring observations [13].
2. Iterative method based on the robust Kalman filter (RKF)
    - Outlier detection: Outlier detection is performed by applying the IQR rule on the standardized residuals after fitting a state-space model to the data.
    - Outlier treatment: An alternative to the state estimator $\hat{\beta}_{t|t}$, inspired by the work by [14] and subsequently by [15], is proposed. In this approach, the state prediction $\hat{\beta}_{t|t}$ is replaced by

$$\hat{\beta}^*_{t|t} = \underset{\beta}{\operatorname{argmin}}\left\{ \left(\hat{\beta}_{t|t-1} - \beta\right)^2 P^{-1}_{t|t-1} + \left(Y^{\text{out}}_t - W_t\beta\right)^2 \sigma_e^{-2} \right\} \tag{3}$$

    where Y^{out}_t is an identified outlier that is replaced by $\hat{Y}^*_t = W_t\hat{\beta}^*_{t|t}$. This proposal considers the robust version of the Kalman filter only at moments at which outliers are detected, as opposed to the original work, in which it is applied at all moments. In the end, the model is iteratively fitted j times to the corrected time series until $\|\hat{\Theta}^{(j)}_{ML} - \hat{\Theta}^{(j-1)}_{ML}\| < \delta$, $j \in \mathbb{N}$, or for some value j.
3. Iterative method based on the Kalman filter for time series with missing values (naKF)
    - Outlier detection: Outlier detection is performed by applying the IQR rule to the standardized residuals after fitting a state-space model to the data.
    - Outlier treatment: Outlier observations Y^{out}_t are assumed to be missing values and the state estimator $\hat{\beta}_{t|t}$ and its mean square error $P^*_{t|t}$ are replaced by $\hat{\beta}^*_{t|t} = \hat{\beta}_{t|t-1}$ and $P^*_{t|t} = P_{t|t-1}$, respectively. The missing observations Y^{out}_t are replaced by $\hat{Y}^*_t = W_t\hat{\beta}^*_{t|t}$ and the state-space model is fitted j times to the corrected time series until $\|\hat{\Theta}^{(j)}_{ML} - \hat{\Theta}^{(j-1)}_{ML}\| < \delta$, $j \in \mathbb{N}$, or for some value j.

The aim of this paper is to investigate under which conditions the presence of outliers affects the estimation of parameters and states in the state-space model and to propose competitive approaches for outlier detection and treatment. Thus, we simulate time series of size n ($n = 50{,}500$), considering for all simulation studies the local level model, which is a simple and particular case of the state-space model (2)–(4), where $W_t = 1$, $\forall t$ and $\phi = 1$, which will be used to illustrate the non-stationary case. The local level model is given by:

$$\begin{aligned} Y_t &= \beta_t + e_t \\ \beta_t &= \beta_{t-1} + \varepsilon_t \end{aligned} \tag{4}$$

In the literature, some approaches have been proposed for the initialization of the Kalman filter for non-stationary stochastic processes. Perhaps the best known is the diffuse initialization ([16]). In this paper, we will use the approximate diffuse initialization, assuming a zero mean and a very large variance of the state ($\sigma_e^2 \times 10^7$).

This study examines two distinct situations: one characterized by non-contaminated data, i.e., the clean data where $e_t \sim N(0, \sigma_e^2)$; $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, and the other involving data that has been contaminated at a rate of $p = 0.05$, i.e., $e_t \sim (1-p)N(0, \sigma_e^2) + pN(10\sigma_e, \sigma_e^2)$; $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$.

For each of the scenarios, the simulation design was formulated with a sample sizes of n = (50,500), and σ_ε^2 and σ_e^2 (0.10, 1.00, 0.05). For each parameter combination, 1000 replicates with valid estimates were considered, i.e., $\sigma_\varepsilon > 0$, and $\sigma_e > 0$; It was considered as convergence criteria $\|\hat{\Theta}_{ML}^{(j)} - \hat{\Theta}_{ML}^{(j-1)}\| < 10^{-4}$ or until $j = 100$. To initialize the Kalman filter, $\mu_1 = 0$ and $P_1 = \sigma_e^2 \times 10^7$ was taken.

To evaluate the quality of the parameter estimates and the k-steps ahead forecasts, it was considered that

- $\text{RMSE}(\Theta) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\Theta_i - \widehat{\Theta}_i\right)^2}$;
- $\text{MAE}(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left|\Theta_i - \widehat{\Theta}_i\right|$.

To evaluate the rate of true outliers detected, two rates were used rate 1 = A/B; rate 2 = A/C, where A is the number of true outliers detected, B is the total number of outliers detected by the method (total number of true and false outliers), and C is the total number of true outliers.

3. Results

In this section, the results obtained from the proposed methodologies are presented. The results of the simulation study are represented in the first subsection. In the second subsection, the application of outlier detection and treatment methodologies are demonstrated via three illustrative examples.

3.1. Simulation Results

Tables 1 and 2 show the RMSE and MAE of the local level model parameters and the one-step ahead forecasts for sample sizes $n = 50$ and $n = 500$ for the simulation study, respectively. In most scenarios, the methodologies improved the accuracy of model parameters and one-step ahead forecasts. However, this improvement was minimal. In fact, there are scenarios where the RMSE and MAE evaluation measures are lower in the non-treated case compared to when outliers are treated; for example, for the scenario $n = 500$, $\sigma_\varepsilon^2 = 0.10$ and $\sigma_e^2 = 0.05$. In particular, LI performed least favourably in comparison to RKF and naKF, especially to estimate the variance of the observation error σ_e^2. For example, for $n = 500$, $\sigma_\varepsilon^2 = 0.10$ and $\sigma_e^2 = 1.00$, in the case of treating outliers by LI, the RMSE of σ_e^2 was 2.0559, while for RKF it was 0.2428 and for naKF it was 0.1168. Overall, it can also be seen that naKF was the method that showed the better performance to improve the accuracy of the parameters and one-step ahead forecasts, especially for $n = 500$. The proposed methodologies had problems in improving the accuracy of the estimates of the level variance σ_ε^2. Finally, regarding the detection of outliers, it is clearly seen the advantage of identifying outliers over standardized residuals, whose means of rate 1 and 2 were higher.

Table 1. Root-mean-square error (RMSE), mean absolute error (MAE), rate 1, and rate 2 of Θ with 1000 simulations of non-stationary time series of sample sizes $n = 50$, considering Gaussian errors (NC = non-contaminated; C = contaminated; RKF = robust Kalman filter; naKF = Kalman filter for time series with missing values).

Parameters			**RMSE**			**MAE**			**Outlier**	**Mean**	**Mean**
σ_ε^2	σ_e^2		σ_ε^2	σ_e^2	$\hat{Y}_{t\|t-1}$ **vs.** Y_t	σ_ε^2	σ_e^2	$\hat{Y}_{t\|t-1}$ **vs.** Y_t	**Detection**	**Rate 1**	**Rate 2**
0.10	0.05	NC	**0.0416**	**0.0276**	**0.4271**	**0.0335**	**0.0217**	**0.3399**		-	-
		C	0.0621	0.2614	0.5243	0.0475	0.2214	0.4033		-	-
		LI	0.0584	0.1772	0.4910	0.0438	0.1286	0.3781	Time series	84%	42%
		RKF	0.0665	0.0910	0.4910	0.0456	0.0718	0.3781	Standardized residuals	74%	88%
		naKF	0.0536	0.0556	0.4667	0.0393	0.0337	0.3607			
1.00	0.10	NC	**0.3114**	**0.1453**	**1.0734**	**0.2488**	**0.1088**	**0.8539**		-	-
		C	0.4638	0.6275	1.2216	0.3644	0.4951	0.9507		-	-
		LI	0.4255	0.5723	1.2127	0.3432	0.4499	0.9421	Time series	45%	8%
		RKF	0.4216	0.4347	1.2048	0.3384	0.3422	0.9387	Standardized residuals	61%	42%
		naKF	0.4285	0.3821	1.2210	0.3422	0.2706	0.9383			
0.10	1.00	NC	**0.0840**	**0.2456**	1.1675	**0.0618**	**0.1977**	0.9326		-	-
		C	14.5332	468.2479	1.4690	1.3638	77.8606	1.1298		-	-
		LI	0.1025	0.3266	**1.1653**	0.0719	0.2373	**0.9250**	Time series	91%	99%
		RKF	0.3768	0.5958	1.2860	0.1245	0.3587	0.9876	Standardized residuals	78%	98%
		naKF	0.4510	0.3155	1.2844	0.1582	0.2525	0.9620			
0.05	0.10	NC	**0.0275**	**0.0329**	**0.4413**	**0.0212**	**0.0260**	**0.3517**		-	-
		C	0.0564	0.4242	0.5416	0.0333	0.3516	0.4180		-	-
		LI	0.0343	0.1501	0.4663	0.0237	0.0830	0.3652	Time series	91%	83%
		RKF	0.0586	0.0710	0.4914	0.0327	0.0557	0.3798	Standardized residuals	75%	97%
		naKF	0.0476	0.0391	0.4714	0.0279	0.0294	0.3635			

3.2. Illustrative Examples

In this subsection, a comparative analysis of the proposed outlier detection and treatment methods using the local level model is presented based on three illustrative examples. The aim is to evaluate the performance of the methodologies from a practical point of view, in terms of outlier detection and treatment and validation of the assumptions (normality and independence of residuals). The three time series that present outliers and are used for illustrative purposes are the following:

- TS1: Number of earthquakes per year of magnitude 7.0 or greater, between 1900 and 1998 (Figure 1);
- TS2: Kiewa River at Kiewa, Victoria, Australia, between 1885 and 1954 (Figure 2);
- TS3: Tree: Beyond Burn, Australia. Pencil Pine, between 1028 and 1975 (Figure 3).

The data is available on GitHub (https://github.com/FinYang/tsdl (accessed on 27 June 2023)) in the Time Series Data Library (TSDL), created by Professor Rob Hyndman.

The data was divided into a training sample (80%) and a test sample (20%). TS1 presents one outlier in the training sample corresponding to the year 1943; TS2 presents one outlier in the training sample (1916) and one in the test sample (1955). TS3 presents 18 outliers in the training sample (16 outliers before 1335 and two outliers corresponding to the years 1770 and 1777, respectively) and three outliers in the test sample, namely 1972, 1973 and 1975.

The results of the local level model fit to the three time series are shown in Table 3.

Table 2. Root-mean-square error (RMSE), mean absolute error (MAE), rate 1, and rate 2 of Θ with 1000 simulations of non-stationary time series of sample sizes $n = 500$, considering Gaussian errors (NC = non-contaminated; C = contaminated; RKF = robust Kalman filter; naKF = Kalman filter for time series with missing values).

Parameters			**RMSE**			**MAE**			**Outlier**	**Mean**	**Mean**
σ^2_ε	σ^2_e		σ^2_ε	σ^2_e	$\hat{Y}_{t\|t-1}$ **vs.** Y_t	σ^2_ε	σ^2_e	$\hat{Y}_{t\|t-1}$ **vs.** Y_t	**Detection**	**Rate 1**	**Rate 2**
		NC	**0.0138**	**0.0086**	**0.4315**	**0.0109**	**0.0068**	**0.3443**		-	-
		C	0.0170	0.2228	0.5303	0.0137	0.2187	0.4115		-	-
0.10	0.05	LI	0.0184	0.2156	0.5561	0.0147	0.2112	0.4193	Time series	52%	4%
		RKF	0.0189	0.0696	0.4913	0.0146	0.0684	0.3822	Standardized residuals	77%	91%
		naKF	0.0181	0.0133	0.4656	0.0137	0.0103	0.3613			
		NC	**0.1156**	**0.0524**	**1.0891**	**0.0934**	**0.0419**	**0.8685**		-	-
		C	0.1376	0.4955	1.2374	0.1112	0.4775	0.9679		-	-
1.00	0.10	LI	0.1454	0.4962	1.3117	0.1165	0.4788	0.9915	Time series	19%	1%
		RKF	0.1366	0.3261	1.2226	0.1102	0.3114	0.9550	Standardized residuals	65%	41%
		naKF	0.1643	0.2065	1.2561	0.1272	0.1803	0.9634			
		NC	**0.0235**	**0.0771**	**1.1685**	**0.0188**	**0.0610**	**0.9324**		-	-
		C	0.0351	4.7013	1.4334	0.0275	4.6231	1.1320		-	-
0.10	1.00	LI	0.0341	2.0559	1.2754	0.0242	1.3978	1.0019	Time series	94%	68%
		RKF	0.0299	0.2428	1.2191	0.0227	0.2255	0.9664	Standardized residuals	89%	100%
		naKF	0.0423	0.1168	1.1950	0.0254	0.0976	0.9436			
		NC	**0.0086**	**0.0100**	**0.4466**	**0.0068**	**0.0079**	**0.3561**		-	-
		C	0.0125	0.4614	0.5647	0.0098	0.4517	0.4417		-	-
0.05	0.10	LI	0.0119	0.3605	0.5628	0.0094	0.3348	0.4290	Time series	81%	23%
		RKF	0.0116	0.0722	0.4893	0.0088	0.0702	0.3854	Standardized residuals	84%	99%
		naKF	0.0104	0.0129	0.4617	0.0077	0.0106	0.3644			

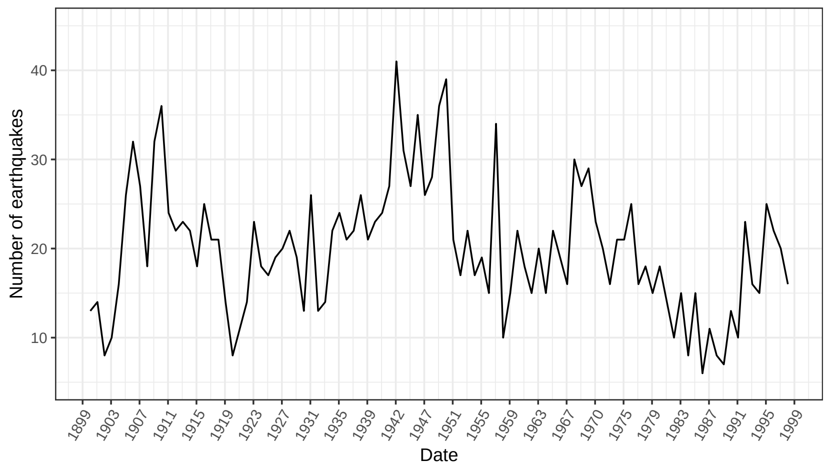

Figure 1. Number of earthquakes per year of magnitude 7.0 or greater, between 1900 and 1998 (TS1).

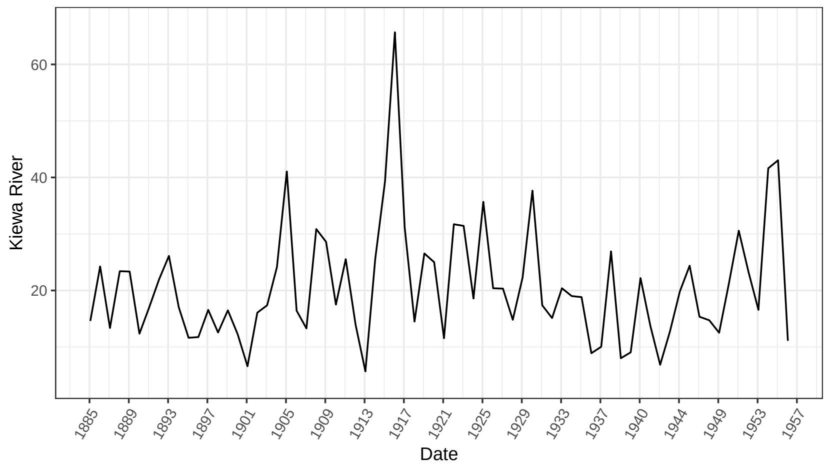

Figure 2. Kiewa River at Kiewa, Victoria, Australia, between 1885 and 1954 (TS2).

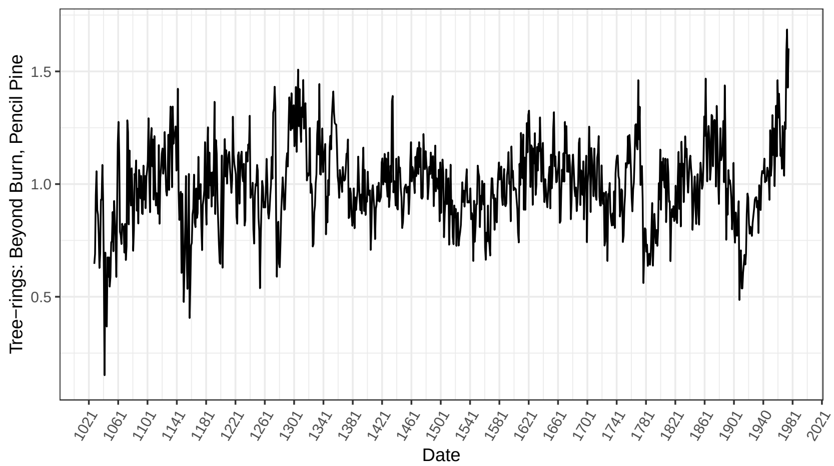

Figure 3. Tree: Beyond Burn, Australia. Pencil Pine, between 1028 and 1975 (TS3).

Table 3. Parameter estimates and respective standard errors of the non-stationary state-space model (local level model); LI—linear interpolation; RKF—robustified Kalman filter; naKF—Kalman filter for time series with missing values.

		σ_ε		σ_e		**log *L***
		Estimate	**(SE)**	**Estimate**	**(SE)**	
TS1	Non-treated	2.7103	(0.6932)	4.8341	(0.5760)	−192.8515
	LI	2.6438	(0.6735)	4.6330	(0.5578)	−190.1958
	RKF	2.9174	(0.6983)	4.0890	(0.5653)	−185.7237
	naKF	3.0671	(0.7237)	3.8387	(0.5844)	−183.6041
TS2	Non-treated	1.6446	(0.8822)	9.3662	(0.9774)	−170.7793
	LI	1.2913	(0.7006)	7.8502	(0.8092)	−161.2743
	RKF	1.1704	(0.6859)	7.7522	(0.7959)	−160.3136
	naKF	1.0999	(0.6905)	7.7692	(0.7967)	−158.2455
TS3	Non-treated	0.0623	(0.0058)	0.1054	(0.0046)	1096.8770
	LI	0.0597	(0.0057)	0.0971	(0.0045)	1149.8800
	RKF	0.0614	(0.0055)	0.1000	(0.0044)	1129.9350
	naKF	0.0601	(0.0055)	0.1020	(0.0044)	1124.1500

After fitting the model to the non-treated data, outliers were detected in the standardized residuals, and these outliers were treated in the two iterative methods, RKF and naKF.

In TS1, two outliers were detected (1943 and 1957). In example TS2, the detected outlier initially remained (1916). Finally, in TS3, where eighteen outliers were initially detected, after the adjustment the residuals showed eight outliers, of which three (1042, 1158 and 1777) were initially detected in the time series.

Table 4 shows the observed evaluation measures and predicted values in the test sample. This table highlights the lowest RMSE and MAE values, with the naKF method performing best. However, the difference between these values is minimal, especially in the case of TS3; therefore, these results are in line with those obtained in the simulation study.

Table 4. Root-mean-square error (RMSE) and mean absolute error (MAE) between the observed and forecasted values via the local level model in the test sample.

		Non-Treated		LI		RKF		naKF	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
TS1	Y_t vs. $\widehat{Y}_{t+k\|t}$	7.0245	6.0496	7.0087	6.0353	6.8205	5.8609	**6.7342**	**5.7788**
	Percentage reduction	-	-	0.22%	4.14%	2.90%	3.12%	**4.13%**	**4.48%**
TS2	Y_t vs. $\widehat{Y}_{t+k\|t}$	11.4091	8.1459	11.3624	8.1456	11.2833	**8.1455**	**11.2249**	**8.1455**
	Percentage reduction	-	-	0.41%	0.004%	1.10%	**0.01%**	**1.61%**	**0.01%**
TS3	Y_t vs. $\widehat{Y}_{t+k\|t}$	0.3759	0.3231	0.3757	0.3229	0.3756	0.3228	**0.3742**	**0.3213**
	Percentage reduction	-	-	0.05%	0.06%	0.08%	0.09%	**0.45%**	**0.56%**

Figures 4–6 show TS1, TS2 and TS3 in black, respectively, the forecasts in red, and the 95% prediction intervals using naKF for the treatment of outliers. The amplitude of the prediction intervals for TS1 (Figure 4) and TS3 (Figure 6) show a considerable increase over time, whereas for TS2 (Figure 5) this increase is minimal, and the interval does not cover all the observations in the test sample.

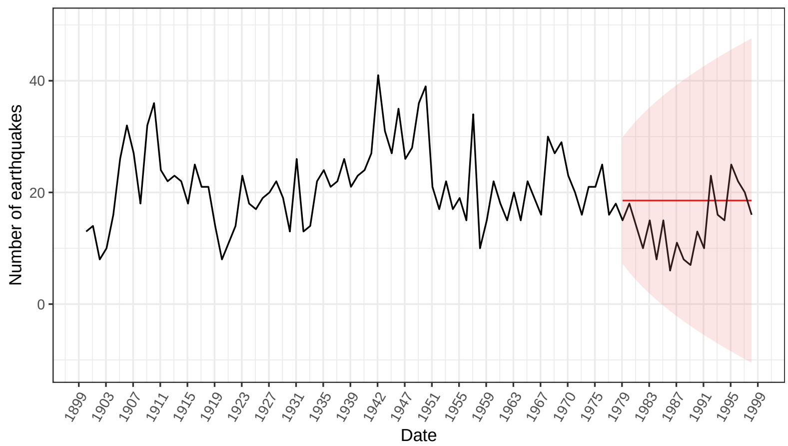

Figure 4. TS1 (black), the k-steps ahead forecasts (red) and the 95% prediction intervals using naKF (red shadow).

Regarding the analysis of the model assumptions, the residuals should behave similarly to white noise. Normality was verified for all models and for all time series: Kolmogorov–Smirnov p values between 0.398 (RKF and TS2) and 0.967 (RKF and TS1). The models for TS1 and TS2 verified the independence assumption: p values ranging between 0.314 (non-treated and TS1) and 0.574 (NA and TS1) from the Ljung–Box test. However, this assumption was not verified for TS3 (all p values of the Ljung–Box test were less than 0.003).

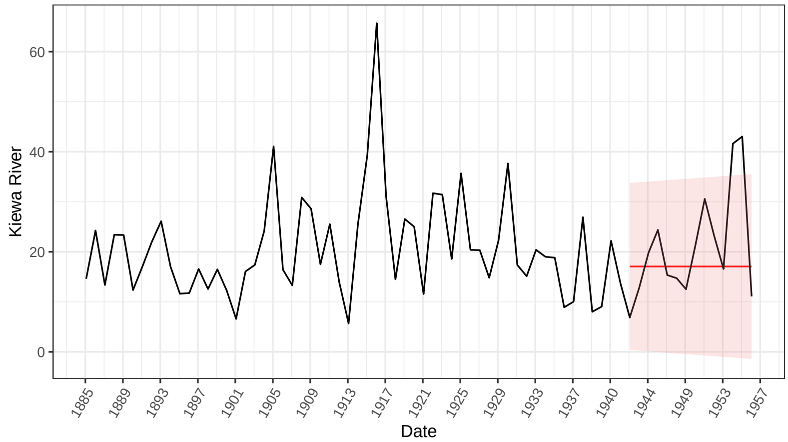

Figure 5. TS2 (black), the *k*-steps ahead forecasts (red) and 95% prediction intervals using naKF (red shadow).

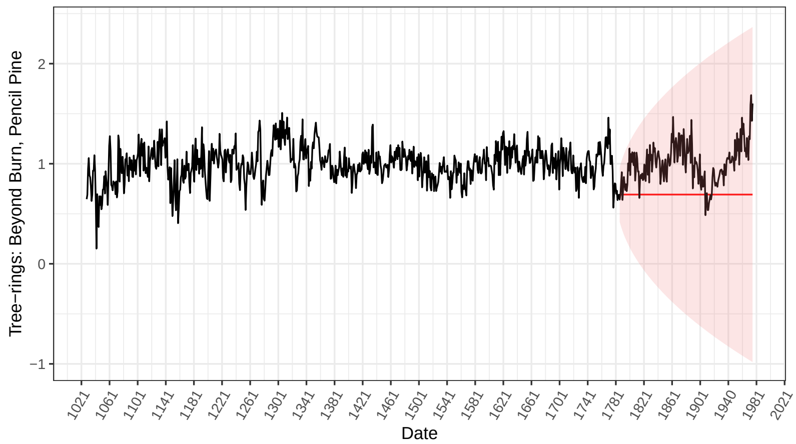

Figure 6. TS3 (black), the *k*-steps ahead forecasts (red) and 95% prediction intervals using naKF (red shadow).

4. Discussion

In this work, three methods for detecting and treating outliers in time series were proposed. This study highlighted the problem of contaminated non-stationary time series from a state-space modelling perspective. To study the impact of outliers on parameter estimates and the observation forecasts, and to make a comparative analysis of the proposed methods, a simulation study was conducted with sample sizes of 50 and 500 with various combinations of parameters, generated using a non-stationary local level model. The data were contaminated at a 5% error rate of the observations. It was found that the proposed methods overall improved the accuracy of the parameters and forecasts; however, this improvement was minimal compared to the contaminated data. The treatment of outliers by naKF and RKF were found to be the most favourable, therefore highlighting the performance of naKF. LI was overall performed the worse. These proposed methodologies were applied to three real time series, where the same conclusion was drawn. In other words, in view of the study's results, the state-space models are generally sufficiently robust, given that they are able to handle non-stationary time series and outliers in a satisfactory way.

Author Contributions: F.C.P., A.M.G. and M.C. contributed to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available in GitHub (https://github.com/FinYang/tsdl (accessed on 27 June 2023)).

Acknowledgments: F. Catarina Pereira was funded by national funds through FCT (Fundação para a Ciência e a Tecnologia) through the individual PhD research grant UI/BD/150967/2021 of CMAT-UM. A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT within the Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM. Marco Costa was partially supported by The Center for Research and Development in Mathematics and Applications (CIDMA-UA) through the Portuguese Foundation for Science and Technology—FCT, references UIDB/04106/2020 and UIDP/04106/2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Triantafyllopoulos, K. The State Space Model in Finance. In *Bayesian Inference of State Space Models*; Springer Texts in Statistics; Springer: Cham, Switzerland, 2021. [CrossRef]
2. Auger-Methe, M.; Newman, K.; Cole, D.; Empacher, F.; Gryba, R.; King, A.A.; Leos-Barajas, V.; Flemming, J.M.; Nielsen, A.; Petris, G.; et al.A guide to state–space modeling of ecological time series. *Ecol. Monogr.* **2021**, *91*, 1–38. [CrossRef]
3. Wu, H.; Matteson, D.; Wells, M. Interpretable Latent Variables in Deep State Space Models. *arXiv* **2022**, arXiv:2203.02057
4. Matsuura, K. Time Series Data Analysis with State Space Model. In *Bayesian Statistical Modeling with Stan, R, and Python*; Springer: Singapore, 2022. [CrossRef]
5. Monteiro, M.; Costa, M. Change Point Detection by State Space Modeling of Long-Term Air Temperature Series in Europe. *Stats* **2023**, *6*, 113–130. [CrossRef]
6. Pereira, F.C.; Gonçalves, A.M.; Costa, M. Short-term forecast improvement of maximum temperature by state-space model approach: The study case of the TO CHAIR project. *Stoch. Environ. Res. Risk Assess.* **2023**, *37*, 219–231. [CrossRef]
7. Shumway, R.H.; Stoffer, D.S. *Time Series Analysis and its Applications: With R Examples*; Springer: New York, NY, USA, 2017.
8. Kalman, R. A New Approach to Linear Filtering and Prediction Problems. *ASME J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
9. Harvey, A. *Forecasting, Structural Time Series Models and the Kalman Filter*; Cambridge University Press: Cambridge, UK, 1990. [CrossRef]
10. Teunissen, P.J.G.; Khodab, A.; Psychas, D. A generalized Kalman filter with its precision in recursive form when the stochastic model is misspecified. *J. Geod.* **2021**, *95*, 108. [CrossRef]
11. Huang, Y.; Zhang, Y.; Zhao, Y.; Shi, P.; Chambers, J.A. A Novel Outlier-Robust Kalman Filtering Framework Based on Statistical Similarity Measure. *IEEE Trans. Autom. Control* **2021**, *66*, 2677–2692. [CrossRef]
12. Auger-Méthé, M.; Field, C.; Albertsen, C.M.; Derocher, A.E.; Lewis, M.A.; Jonsen, I.D.; Flemming, J.M. State-space models' dirty little secrets: Even simple linear Gaussian models can have estimation problems. *Sci. Rep.* **2016**, *6*, 26677. [CrossRef] [PubMed]
13. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd ed.; OTexts: Melbourne, Australia, 2018.
14. Cipra, T.; Romera, R. Kalman filter with outliers and missing observations. *Test* **1997**, *6*, 379–395. [CrossRef]
15. Crevits, R.; Croux, C. Robust estimation of linear state space models. *Commun. Stat.- Simul. Comput.* **2019**, *48*, 1694–1705. [CrossRef]
16. Durbin, J.; Koopman, S.J. *Time Series Analysis by State Space Methods*, 2nd ed.; Oxford Statistical Science Series; Oxford University Press: Oxford, UK, 2013. [CrossRef]