FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Forecasting Milk Production

Séries Temporais e Previsão

Nádia Soares

January 2019

Contents

1 Abstract

Milk is a vital product in many economies around the world. It is a big part of the agriculture sector and provides exportation and jobs for many communities. Forecasting its production can help planning and managing operations accordingly. This project makes use of ARIMA models and exponential smoothing methods in order to predict the production of this product. A monthly time series of milk production from Canada, collected between 1995 and 2013 was explored, described and used to obtain the forecasts.

2 Introduction

The agriculture sector and its livestock sub-sector are vital industries for several nations. It is a vital part of the global food system and it plays a key role in the sustainability of rural areas in particular.

For certain parts of the world, milk production value accounts for more than 20% of the total agricultural value. It provides a significant amount of exportation and jobs to those communities.

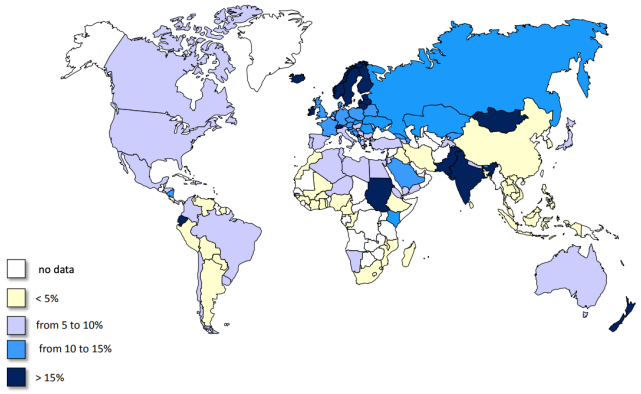

Figure 1: Share of milk in total agricultural production - situation in 2010 according to countries. Taken from [1]

Given all of its value to economies, the forecasting of milk production becomes increasingly relevant.

Forecasting plays a pivotal role in the operations of modern management. It is an important and necessary aid to planning and is the backbone of effective operations. And together with demand forecasting, production forecasting facilitates critical business activities like budgeting, financial planning, sales and marketing plans, pricing planning, risk assessment and formulating mitigation plans.

3 Objective

The goal of this project is to explore a milk production dataset, construct models that explain its correlation and forecast the series future values.

Different methods will be tested and compared, namely, ARIMA and exponential smoothing methods. The steps of the Box-Jenkins approach are followed in order to estimate a good model and obtain the forecasts.

We started by presenting an intuition for the need of forecasting production and daily products, like milk, in particular as well as the objectives of this work. Then, the dataset is presented and an exploratory data analysis is performed. We detail the transformation applied to the time series in order to remove its heteroscedasticity and the detrend and deseason processes. An analysis of the residuals is performed. We move on to estimating and applying models using ARIMA and exponential smoothing. We forecast new values using the testset and the results of the different methods are shown and compared. Finally, we draw conclusions from this project.

4 The dataset

The dataset used for this project was obtained from a github repository: https://raw.githubusercontent.com/ricardoscr/
Data-Science-Certificate/master/02-Methods/CADairyProduction.csv

The csv contains several time series for different dairy products: cotagecheese, icecream and milk. This report focus on the milk production. The data come from Canada.

The time series contains 228 monthly observations, starting in January, 1995 and ending in December, 2003. The values are presented in pounds per cow.

4.1 Exploratory data analysis

What follows is an exploratory analysis done to better understand the data at hand.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1995	2.112	1.932	2.162	2.130	2.227	2.124	2.184	2.152	2.062	2.121	2.030	2.091

Table 1: First 1 year of the time series in pounds per cow.

Table 1 shows the first 6 values of the time series (from January to June, 1995).

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.932	2.583	3.071	2.947	3.344	3.804

Table 2: Statistical description of the time series.

Some statistics of the time series are presented in Table 2.

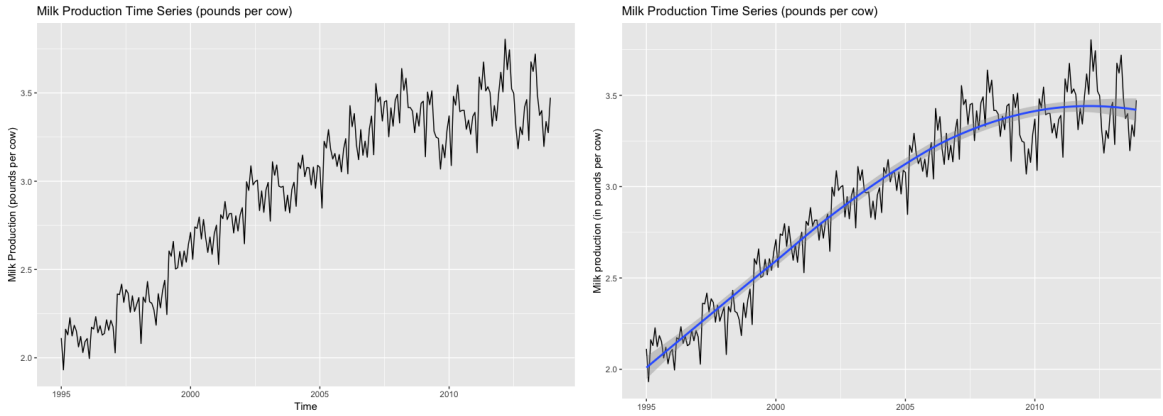

Figure 2: a) Values of milk in pounds per cow plotted over time. b) A smooth trend line was added for visualization purposes.

By analyzing the time series plot shown in Figure 2, we can see that the series presents an upward trend and cycles that seem to be season effects. Also, there is heteroscedasticity: the variance seems to be increasing over time.

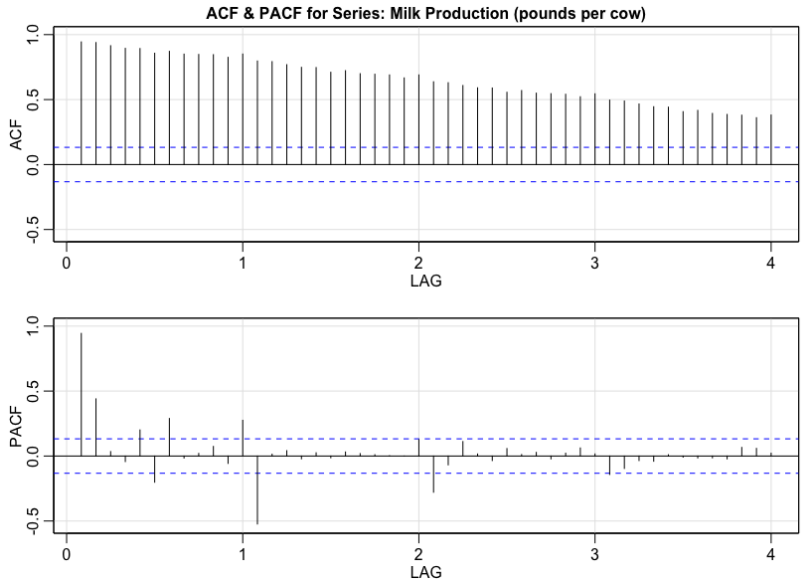

Figure 3: Autocorrelation and partial autocorrelation functions of the milk production series

The autocorrelation and partial autocorrelation functions shown in Figure 3 also indicates the presence of cycles since correlation does not decay at a constant rate when the lag increases. It even increases from lag 0.5 to 0.6, 0.9 to lag 1 and on each 0.5 increase in lag. So there is a strong correlation between the values one year apart, meaning this cycles are actually seasonality.

The ACF and PACF indicate that this is not white noise, but a series of correlated values. Nevertheless, we can do a Ljung-Box test on the time series to formally validate this claim. The p-value is very low so we reject the null hypothesis of absence of serial correlation. The small p-value is significant at p <0.001 so this supports our ACF plot consideration above where we stated it is likely this is not purely white noise and that some time series information exists in this data.

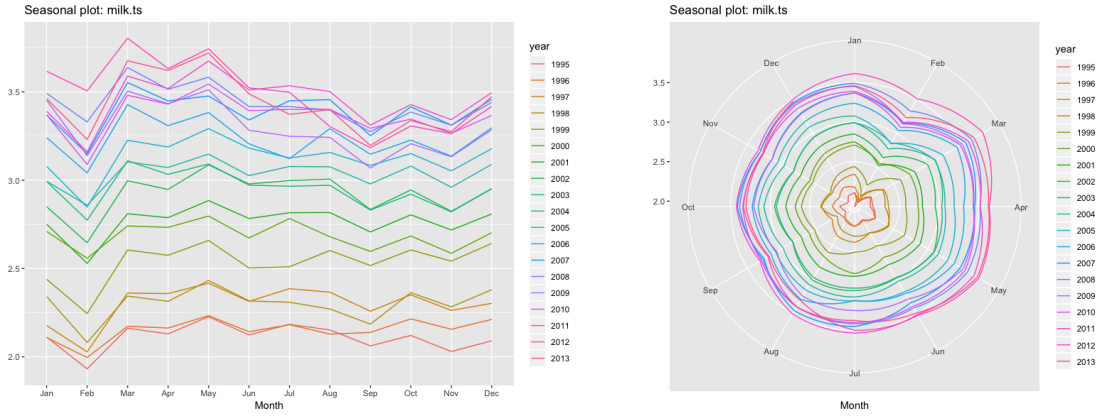

Figure 4: Season plots a) linear and b) polar. Both plots illustrate a sharp decrease in values in February, September and November and picks in March and May.

The seasonal plot in figure 4 allows the underlying seasonal pattern, and can be useful in identifying years in which the pattern changes. Here, we see that milk production has consistently increased over the years as the lines representing earlier years are lower in the plot, and the lines representing recent years are higher. It is also noticeable that, in the last few years, variance has increased and the months of low production of the last years are even lower than in previous years.

Also, we see that milk production tends to be the lowest in some months like February, September and November and typically peaks in March or May.

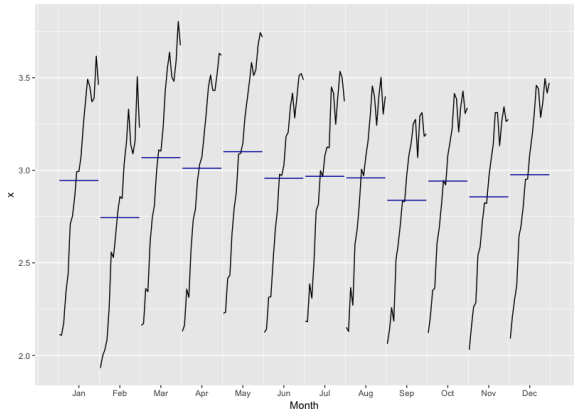

Figure 5: Month plot. Each month is a subseries

The same patterns can be seen in figure 5 along with the monthly variations and means, in blue.

5 Transformations & Decomposition

5.1 Removing Heteroscedasticity

Non-constant variance - heteroscedasticity - can cause problems in the models used. We will need to correct it before modeling to remove the increase in variance in our series. We will apply the log function to the time series. If the series is multiplicative ($Yt = Season_t * Trend_t * E_t$), this will also convert it to an additive time series ($Yt = Season_t + Trend_t + E_t$).

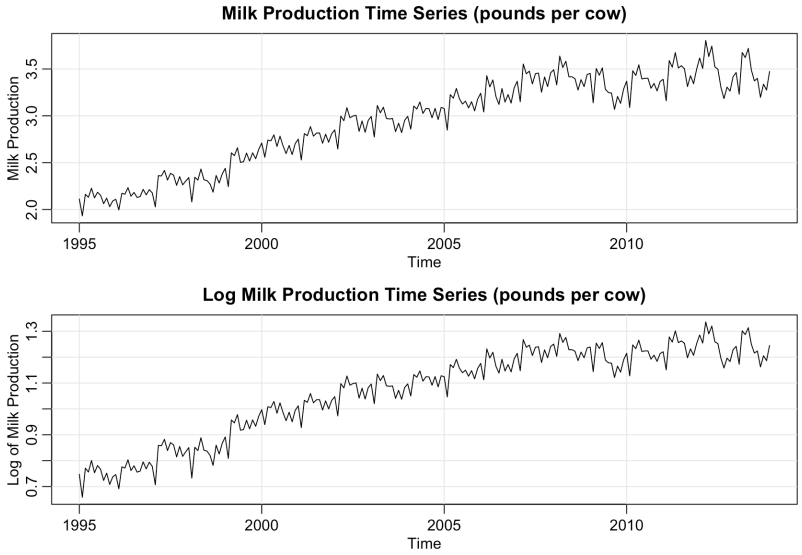

Figure 6: a) original and b) logged milk production time series

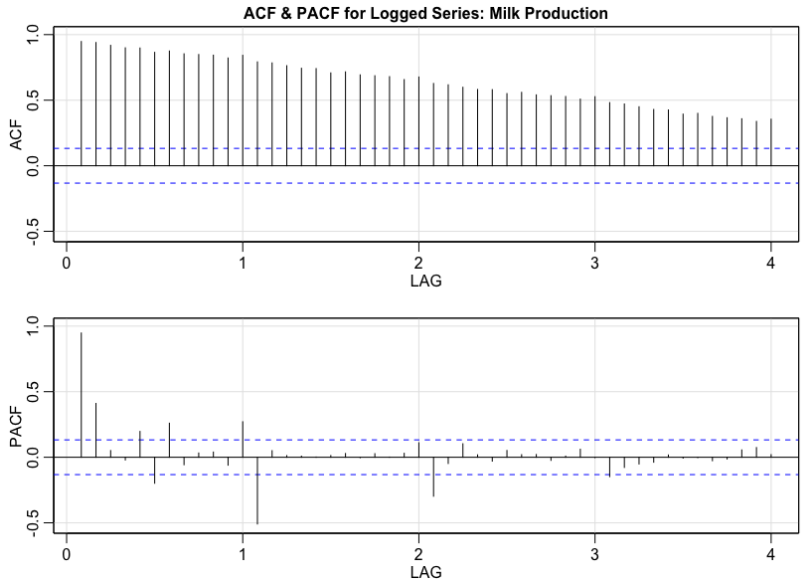

Figure 7: Result of the Box-Ljung test performed on the first 4 lags

Log serves our purpose since it did remove the increase in variance, but we could also use the more general Box-Cox transformation (of which log is a special case).

The optimal value of the lambda Box-Cox parameter is -0.2210074. And the resulting series is very similar to the logged series.

We use the logged series moving forward.

5.2 Removing Trend

To remove the trend we can fit a linear model.

```
fit <- lm(milkLog~time(milkLog))
```

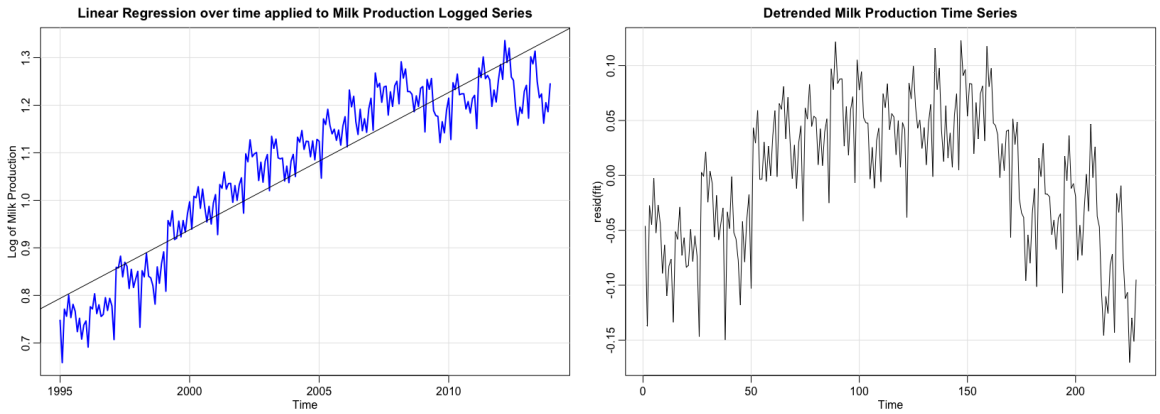

Figure 8: a) Linear regression over time. b) Time series detrended using the linear model above.

The linear model that is applied to the time series is shown in figure 8 a). b) shows the resulting detrended series. It is clear that, due to the trend not being linear, the linear model is not enough to detrend it.

We will need to fit a degree two polynomial model to get a detrended series.

```
fit <- lm(milkLog~poly(time(milkLog), 2)
```

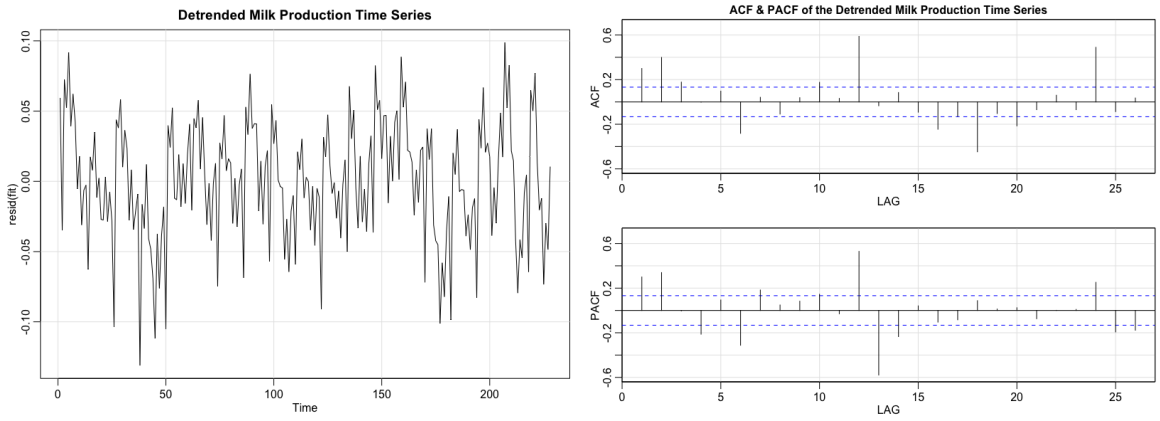

Figure 9: a) Detrended time series using a degree two polynomial and b) its ACF and PACF functions

The ACF of the detrended series looks cyclic. Lags 12 and 24 show great correlation and lags 6 and 18 shows the minimum correlation of the first and second years, respectively. This means there must be seasonality in the time series.

5.3 Removing Seasonality

In order to remove this seasonal effect, we decomposed the series using the STL method and seasonally adjusted, figures 10 and 11.

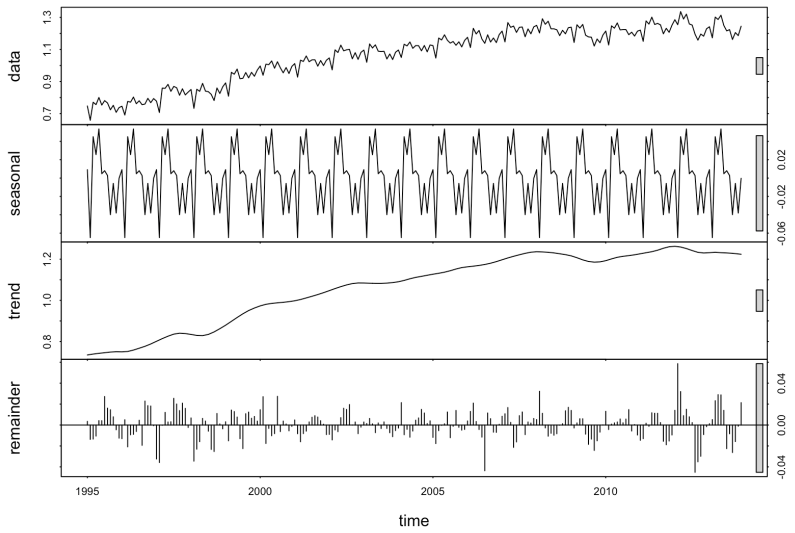

Figure 10: Series decomposition using STL. The data and its three components are depicted.

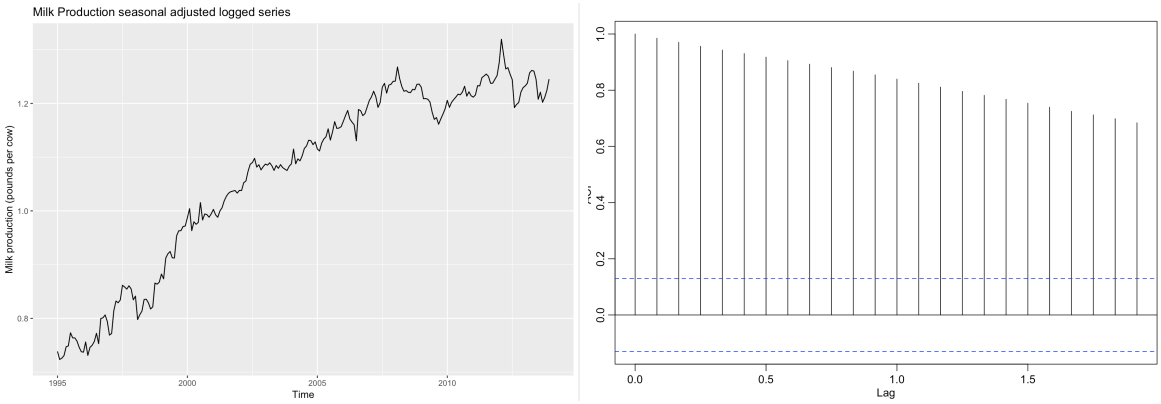

Figure 11: a) Seasonally adjusted time series and b) its ACF

The season plot in figure 12 is now much smoother.

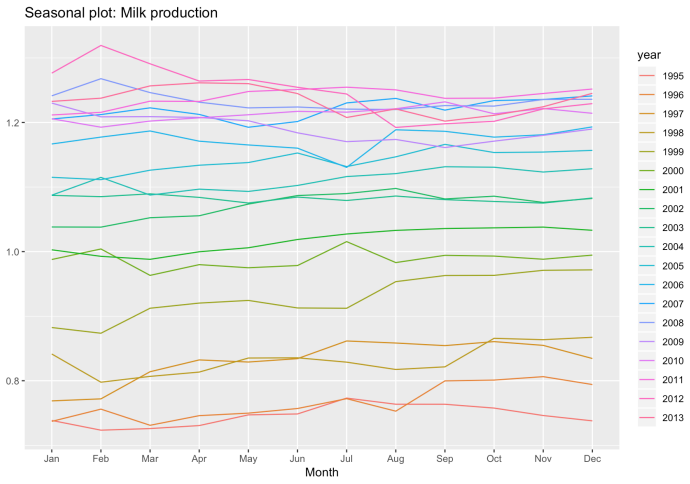

Figure 12: Season plot of the adjusted time series

Applying the polynomial regression above to the seasonally adjusted series removes its trend, figure 13.

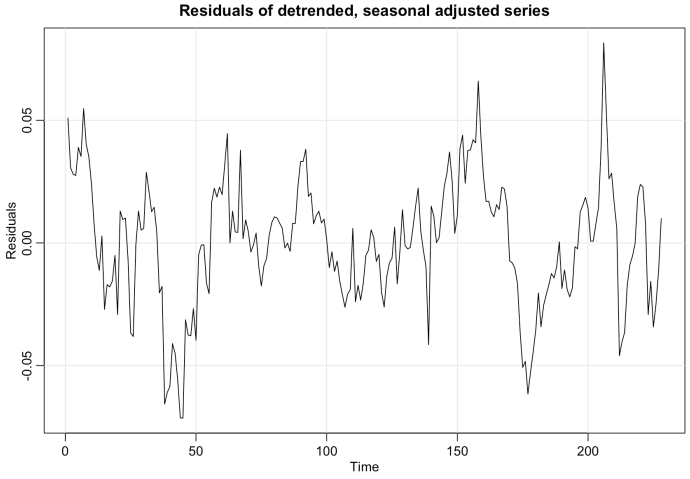

Figure 13: Series of the removed trend and season.

Analyzing the ACF of the remainder component of the decomposition, figure 14, it is visible that there is still some unexplained correlation. We try to fit it using an AR(1) model and afterwards, most of the correlation is explained.

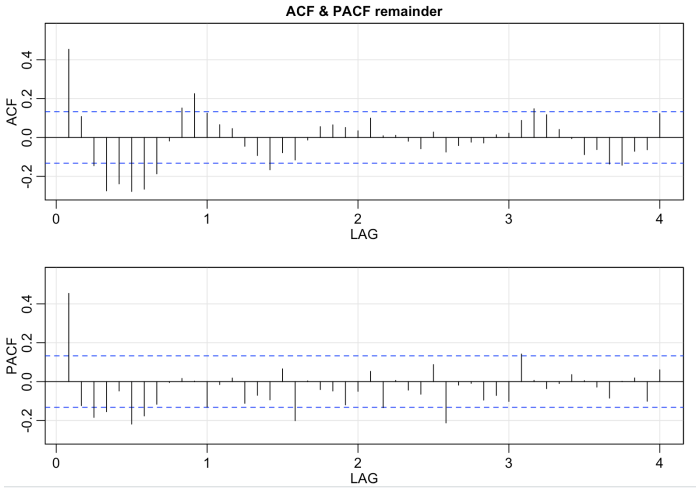

Figure 14: ACF of the remainder from the STL decomposition

6 Modelling

The first step in order to estimate the models is to divide the series in train and test sets. The first 18 years, from 1995 to 2012, form the training set while the last one year, 2013, was used as the testing set.

6.1 Modelling using ARIMA

Differentiating the time series

To verify if the time series is stationary an augmented Dickey-Fuller (ADF) test will be used.

This types of tests are called unit root tests as the null hypotheses tests the presence of a unit root. So the alternative hypothesis is the model being stationary.

A visual inspection of the time (logged) series tells us that it is non-stationary because the mean varies over time. The ADF test confirms it by having a p-value of 0.2702, thus not rejecting the null hypothesis.

Using the ndiffs function we can also perform unit root tests (KPSS, Augmented Dickey-Fuller or Phillips-Perron) and they suggest the number of differences to apply to the data. In this case, they all return 1 difference.

So lets apply the first difference to the data.

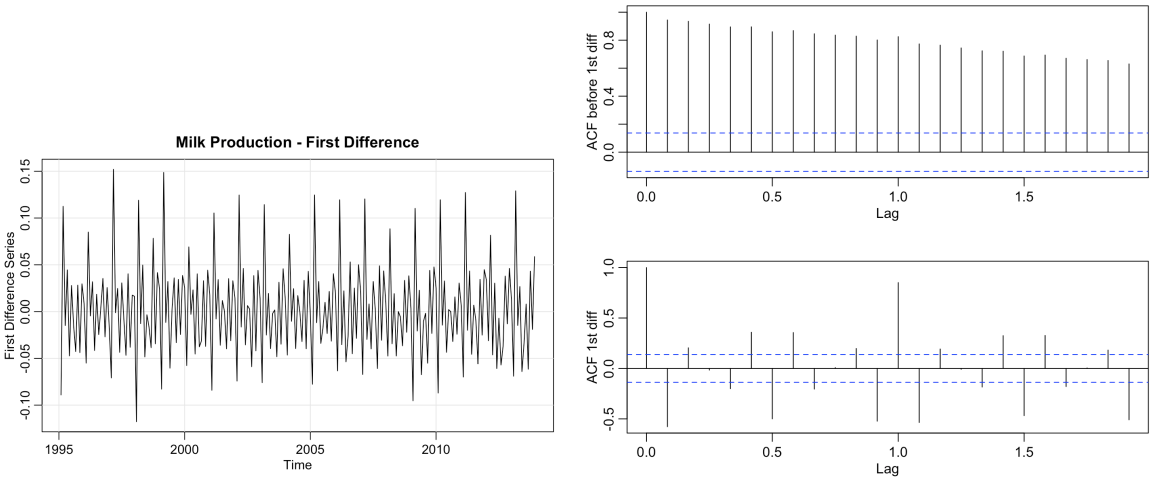

Figure 15: a) First difference of the logged time series and b) its ACF and PACF

The first difference did decrease the variance from 0.02913752 to 0.00245081, and it removed the trend.

The season component of the series is clear on the ACF's.

For the seasonal difference there is also unit root tests that suggest the number of differences to apply. In this case, the nsdiffs function, using the default test, seas, by Wang, Smith and Hyndman, return 1 seasonal difference.

Applying the 12 difference we obtain the results in figure 16.

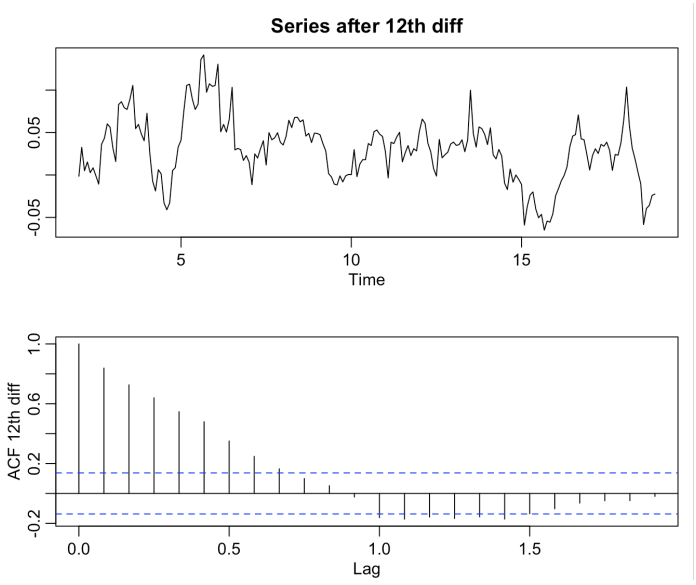

Figure 16: a) 12th difference and b) its ACF

The variance is now 0.001460643.
Finally we try both differences and the variance decreases to 0.0004554067. Results in figure 17.

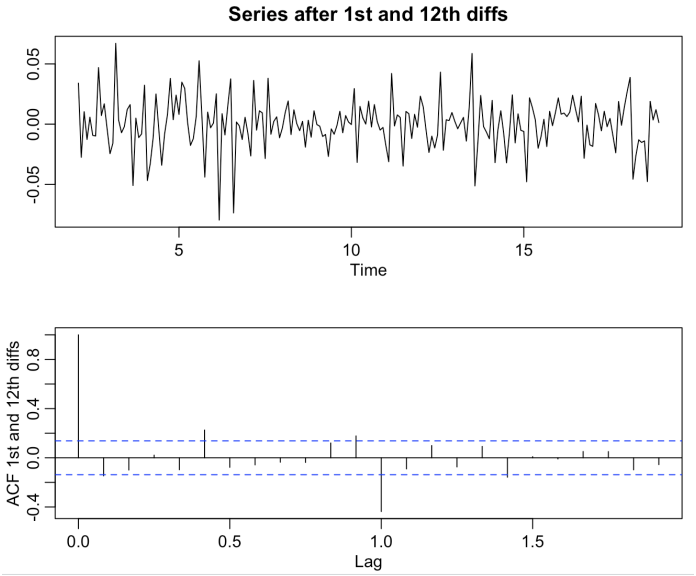

Figure 17: a) Both 1st and 12th differences and b) the ACF of the result

Now, the ADF test has a small p-value and we reject the null hypothesis, estimating the series is indeed stationary.

Estimating orders

We will now estimate the orders of the model.

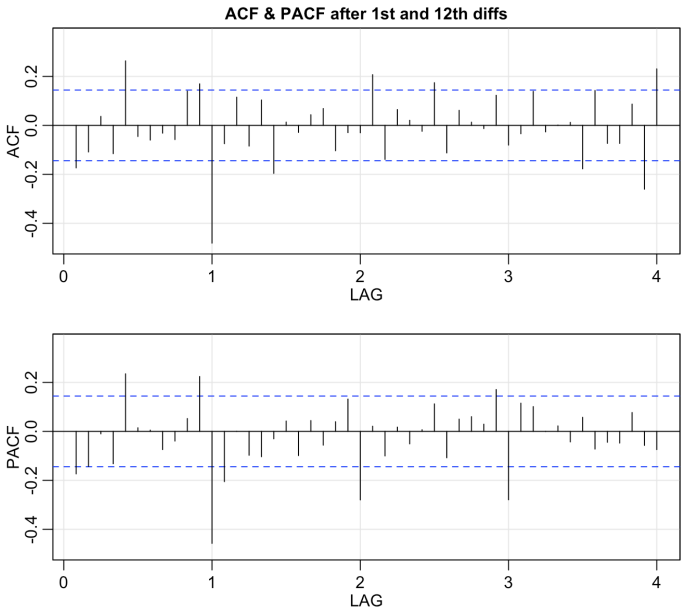

Figure 18: ACF and PACF of the differenciated series

Looking at seasonal part of the ACF PACF in figure 18, we can see that there are high correlation at year one on the ACF, and then the other years (2, 3 and 4) are not that significant. The seasonality of the PACF decreases slowly. This indicates an MA(1) model for the seasonal part.

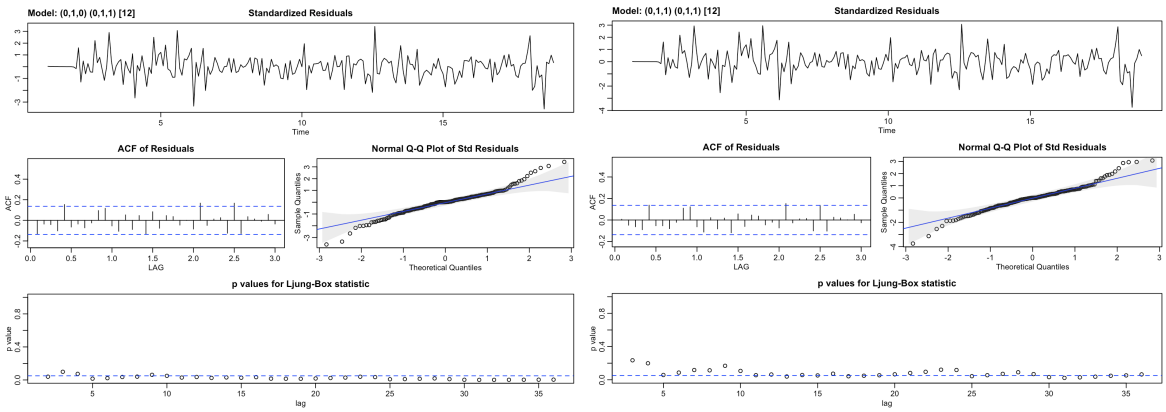

Figure 19: Result applying a a) SARIMA(0,1,0)(0,1,1)[12] and a b) SARIMA(0,1,1)(0,1,1)[12]

We apply a SARIMA(0,1,0)(0,1,1)[12], figure 19 a), and the seasonal effect seems to be explained, but the residuals are not normal, the ACF shows some correlation and the p-values are very small indicating that the null hypothesis of the residuals being uncorrelated should be rejected - indicating correlation.

We then tried to module the non seasonal part, by testing some different models and analyzing their residuals and AIC, AICc and BIC criteria.

	AIC	AICc	BIC	non-significant coefficients
(0,1,0)(0,1,1)	-7.247317	-7.237797	-8.23169	no
(0,1,1)(0,1,1)	-7.252711	-7.242927	-8.221458	no
(0,1,2)(0,1,1)	-7.246418	-7.236281	-8.199539	yes
(0,1,5)(0,1,1)	-7.241975	-7.230223	-8.148218	yes
(1,1,0)(0,1,1)	-7.251354	-7.241571	-8.220102	yes
(5,1,0)(0,1,1)	-7.24642	-7.234668	-8.152662	yes
(1,1,1)(0,1,1)	-7.252046	-7.241909	-8.205167	yes

Table 3: Criteria of different models tested

On the non-seasonal part, the MA(1) seems to be the best model, figure 19 b). That is also the one chosen by the auto.arima function by minimizing the AIC.

Its residuals are not normal, the ACF shows correlations are mostly non-significant. Most of the p-values in the Ljung-Box test are significant but not by much.This is the best model in terms of AIC and AICc, but the model with only a seasonal MA(1) has a better BIC. Notice that bigger models have good AIC and AICc, but are penalized in the BIC for having non-significant coefficients.

Finally, figure 20, shows the predictions made for the test set using this model.

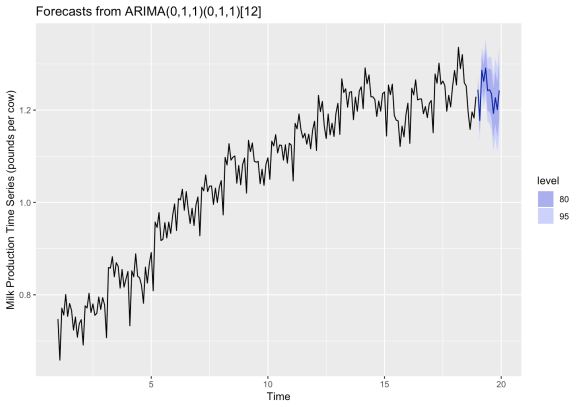

Figure 20: Predictions obtained using the SARIMA(0,1,1)(0,1,1)[12] model

	low	high	Contains	Observed value	Forecast	Percentual Error
1	1.178027	1.236925	FALSE	1.241846	1.207476	0.0276768001
2	1.283547	1.358838	FALSE	1.172792	1.321193	0.1265363867
3	1.257720	1.346424	TRUE	1.301825	1.302072	0.0001898126
4	1.280898	1.381238	TRUE	1.287026	1.331068	0.0342195676
5	1.229445	1.340205	TRUE	1.313724	1.284825	0.0219974934
6	1.229234	1.349515	TRUE	1.249615	1.289374	0.0318171073
7	1.225481	1.354583	FALSE	1.215803	1.290032	0.0610539955
8	1.178588	1.315945	TRUE	1.223775	1.247266	0.0191955117
9	1.208709	1.353853	FALSE	1.162213	1.281281	0.1024496949
10	1.175766	1.328300	TRUE	1.205372	1.252033	0.0387110051
11	1.212616	1.372197	FALSE	1.186318	1.292407	0.0894269064
12	1.224892	1.391225	TRUE	1.245019	1.308059	0.0506338713

Table 4: Confidence interval of the SARIMA predictions

6.2 Modelling using Exponential Smoothing

From the different exponential smoothing methods, it is expected that Holt-Winters has the best results as it accounts for seasonal effects. We compare the forecasting accuracy of different variations using cross-validation.

	MSE	MAE
Additive	0.002234567	0.03180321
Multiplicative	0.002072222	0.03030065
Additive & damped	0.002463025	0.03470061
Multiplicative & damped	0.002567394	0.03536761

Table 5: Values of the mean square error and mean absolute errors for comparing the different variations of the Holt-Winters' seasonal method, using cross-validation

As it is possible to see from table 5, the multiplicative method had better results for both criteria, followed by the additive method. We will compare their AIC's and predictions.

	AIC	AICc	BIC
Additive	-558.1649	-555.0740	-500.7852
Multiplicative	-517.1759	-514.0850	-459.7962

Table 6: AIC, AICc and BIC values for additive and multiplicative Holt-Winters

Looking at the AIC, figure 6, the additive method appears to be a better fit.
The predictions of both methods are shown in figure 21.

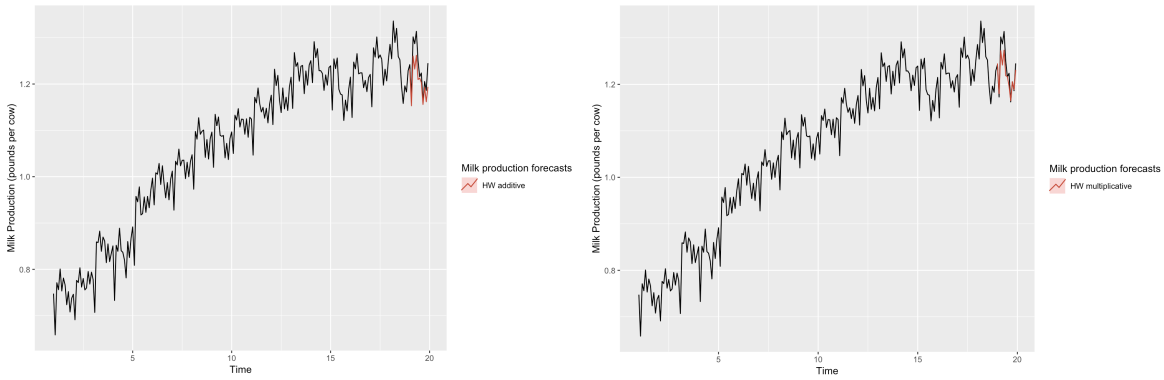

Figure 21: Predictions of the a) additive b) and multiplicative Holt-Winters for the test set

7 Results

As is possible to see in table 7, the SARIMA(0,1,1)(0,1,1)[12] model had the best results. It out performed both Holt-Winters, as all the error metrics evaluated were smaller. It also produced better predictions than the SARIMA model with only seasonal part.

	ME	RMSE	MAE	MPE	MAPE
SARIMA(0,1,0)(0,1,1)[12]	-0.005312542	0.01943295	0.01630804	-0.4800478	1.328801
SARIMA(0,1,1)(0,1,1)[12]	-0.00337877	0.0182781	0.01554326	-0.3215337	1.262382
Additive HW	-0.01790848	0.03329671	0.03021828	-1.426497	2.435279
Multi HW	-0.02024682	0.02862476	0.02523071	-1.626836	2.031836

Table 7: Error metrics of the different models, SARIMA, additive Holt-Winters and multiplicative Holt-Winters

Figure 22 helps comparing the two methods. Note that exponential smoothing methods alone are only capable of generating point predictions, while ARIMA can generate intervals.

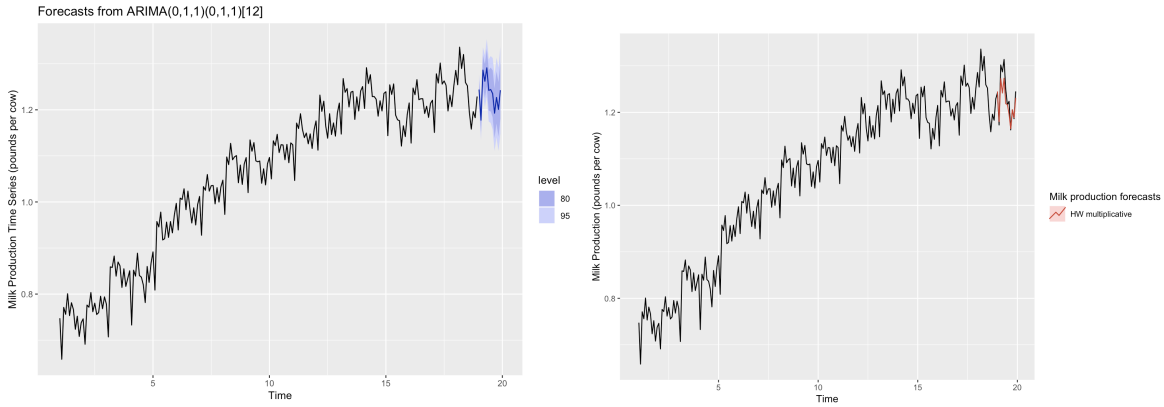

Figure 22: Predictions obtained using the a) SARIMA(0,1,1)(0,1,1)[12] model and the b) multiplicative Holt-Winters.

8 Conclusion

While ARIMA models use the whole time series to create their forecasts, exponential smoothing methods give a bigger weight to most recent observations. In this particular dataset, though, a SARIMA model had the best results.

In the future it would be interesting to experiment with state space models and compare the results obtained using the original series and the log series after doing the inverse transformation.

We could also try to model the dataset variance using a garch model and do a multivariant analysis using the milk data and the ice cream data present in the same dataset.

References

[1] IDF. IDF Factsheet-February 2013 Scientific excellence Industry applicability Strategic networking Global influence Fig. 2: Share of milk in total agricultural production-situation in 2010 according to countries. (February), 2013.

[2] OECD and FAO. Dairy and dairy products. *Agricultural outlook 2018-2027*, page 12, 2018.

[3] B Blaskó. World Importance and Present Tendencies of Dairy Sector. *Production*, 2010.

[4] https://otexts.org/fpp2/.