Time Series Project
year 2012-2013

Analysis of Pollen Allergies Through Google Trends Data

Authors:
Frederic Boua
frederic.boua@epfl.ch
Yoann Trellu
yoann.trellu@epfl.ch

June 7, 2013

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Contents

1 Introduction

We present in this work a time series analysis on the topic of allergy trends within Switzerland. We work here with a quite recent tool in the name of Google Trends. We have at our disposition a set of data of web searches related to "Allergies", as well as measures of pollen concentration for different species in several weather stations from Switzerland. Previous studies on Google Trends [4], motivated our choice for this project.

The objective of our study will be to analyze the time series of search counts to then, with the help of the pollen data, build a linear model to explain the allergies by the pollen. We will assume two things :

1. Google Trends data reflects accurately the population behavior.
2. Allergies due to others sources than pollens are considered white noise, in other words they do not depend on the time of the year. They have a constant mean and a constant variance.

The first question we ask ourselves is what kind of models are suitable for the data. The second one is how should one relate allergies to pollens in a time series framework.

2 Data presentation

The data considered are data from Google Trends [2] from the sub-category "Allergies", the region have been narrowed to Switzerland, the data runs from 2004-01-04 to 2013-03-23 and are averaged on a weekly basis. Categories encompass all searches on Google linked to the category and in this case "Allergies". This gives a set composed of 481 observations, for a time period of more than 9 years. Here are presented the key points of the data collection process and how they are treated by Google:

- The values published by Google are the probability of a random user to search the term in question. This probability is based on a sampling of searches done during the period and in the region selected.
- If the data are insufficient to give coherent result, a zero is inserted.
- For terms search, the number of searches of the word is normalized by the total number of search. This implies for example that a down trend does not necessary represents a diminution of searches but a relative diminution of searches.

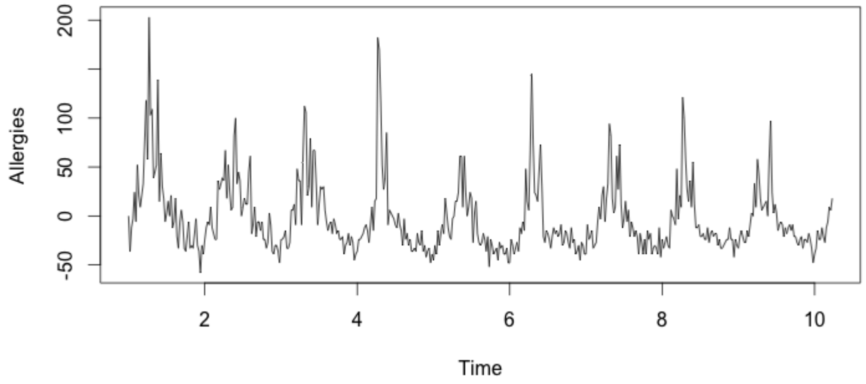

Figure 1: Data of allergies with the percentage of variation on y and the period of 52 weeks on x.

- The scaling applied to categories is slightly different, it shows a percentage of growth in respect to the first value of the series.

In appendix A you can find the exact response of Google to those questions. Figure 1 displays a first plot of the data, this plot shows an obvious annual seasonality as we could expect. The second pertinent observation, is that peaks amplitude vary from year to year.

In this section a first general approach to the data is taken, we plot some basic features of the data and analyze them. The STL decomposition (Figure 2) gives a good summary of the properties of our data: it shows the obvious seasonality, the lack of trend and the remainders where patterns can still be observed which suggests a deeper modeling. When modeling the series, we applied a variance stabilizer in the form of a square root. After this stabilization, the periodogram (Figure 3) as well as the STL decomposition confirm our assumption of the annual seasonality. Indeed in Figure 3, the periodogram with different levels of smoothing, shows a major peak at 1 and a second smaller peak at 2 and 3. The first peak is the annual cycle and the others are the harmonics. Other from that we don't have any signal.

In order to obtain a stationary time series the data have been differenced at lag 52 to remove the seasonality. A differencing at lag one is not necessary since our data show only few to no trend as seen on the STL decomposition. Figure 4 displays the ACF and PACF of the differenced data. It suggests, by their strong peaks on the seasonal component, the fit of a SARIMA model.

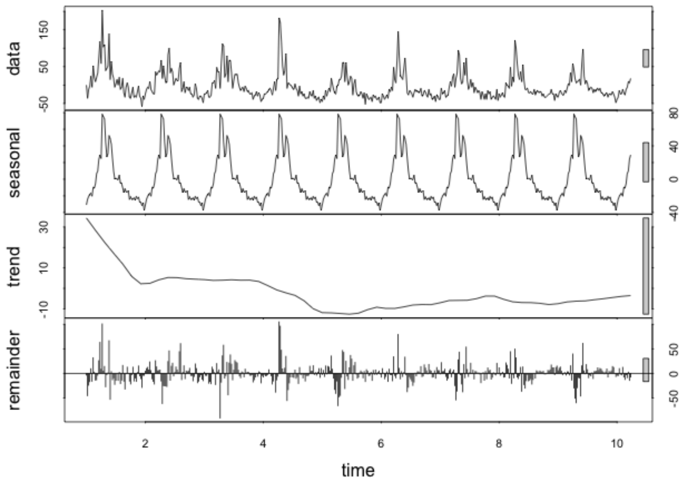

Figure 2: STL decomposition of the data : Seasonal contribution, trend, remainders.

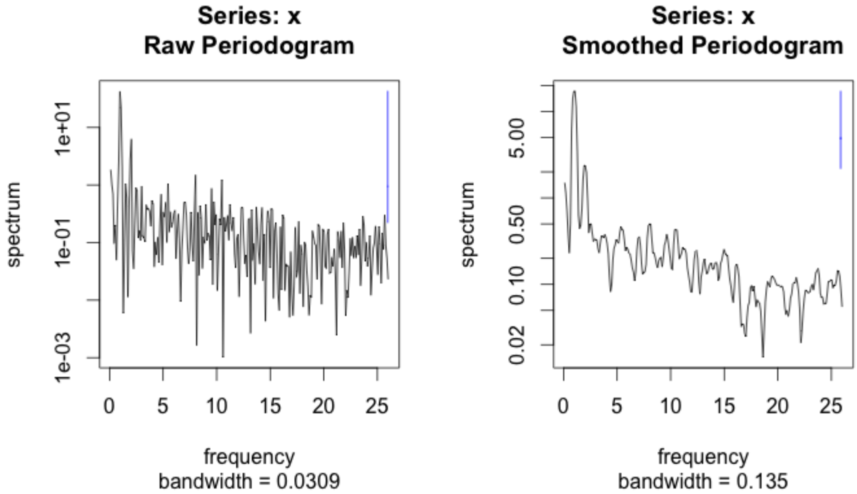

Figure 3: Raw and smoothed periodogram of the data, error bar.

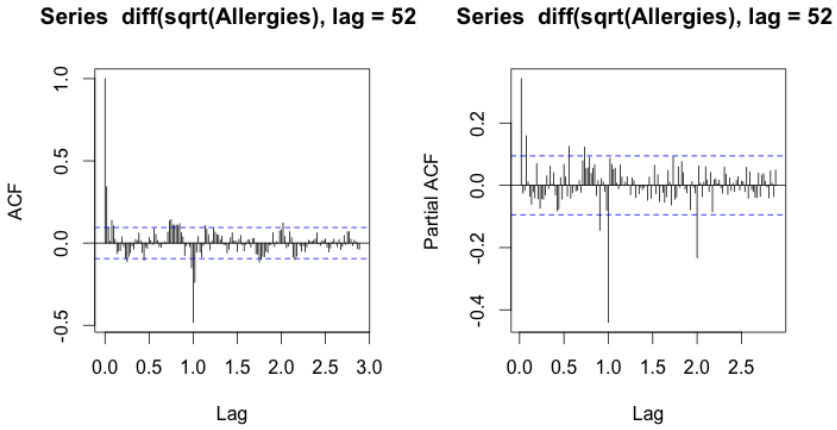

Figure 4: ACF and PACF of the differenced series $lag = 52$, confidence interval for white noise.

Our second set of data is the concentration of pollens in the air. Those data were collected by the Federal Office of meteorology and Climatology MeteoSwiss [1] on a daily basis. The procedure is done by aspiring air and analysing under a microscope the particle collected in a filter during aspiration (see appendix B). The set of data contains the concentration of 7 pollens, responsible for most of the allergies, collected in 11 stations in Switzerland. In order to have a global vision and to minimize the impact of the numerous missing values, we treated the data as follow after aggregating them weekly.

- In order to have weekly values of a specific pollen we averaged the value of the concentration of a given pollen on all the stations (if a value is missing the weekly averaged is performed on the remaining stations).
- In order to have a general vision of all seven pollens in the eleven stations, we summed the values of all pollens in each station, and then averaged this sum on all the stations (a missing value in a pollen observation results in NA in the sum, the average on the station is processed as explain above).

As our goal is to investigate a dependence between the number of allergies search counts and the data on the pollen. To get an idea of the direction we wanted to take, we normalized between 0 and 1 both sets of data and overlapped them on the same graph.

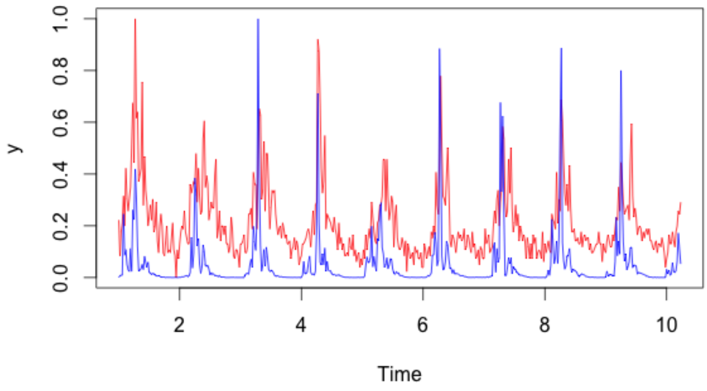

Figure 5: Overlapping of the normalised data, Allergy searches in red, Pollens in blue.

Figure 2 shows an obvious match in frequency and suggested us to try to fit the data on allergies with a linear model combined with an ARMA on the residuals.

3 Model Fitting

3.1 SARIMA Fitting

In this section we try a first model on allergies search counts. The initial data analysis suggested a SARIMA $(p, d, q) \times (P, D, Q)_s$ as defined in the lecture notes [3, p. 87]. In order to determine the different coefficients, we interpreted the results of the first data analysis. First, the periodogram showed a period of 52 weeks, implying $s = 52$. The strong seasonality, the lack of trend as well as the stationarity of the data differenced once at lag 52 suggest values of $D = 1$ and $d = 0$. Finally the analysis of the ACF and PACF (see figure 4) gives information to fix $Q = 1$. To fix the other coefficient as the ACF and PACF analysis gives us uncertain results we decided to compare, based on the loglikelihood and the Akaike Information Criterion (AIC), the different plausible models. Those models were selected by testing reasonable values, based on the estimate of ACF and PACF, for the uncertains coefficients. Figure 6 shows a table with the best results.

$(p,d,q)\times(P,D,Q)_s$	log likelihood	AIC
$(1,0,2)\times(2,1,1)_s$	-1124.81	2263.62
$(2,0,1)\times(1,1,1)_s$	-1130.97	2273.94
$(1,0,1)\times(1,1,1)_s$	-1132.72	2275.44
$(2,0,2)\times(0,1,1)_s$	-1132.77	2277.54
$\vdots$	$\vdots$	$\vdots$

Figure 6: Coefficient of SARIMA models .

Based on the log likelihood and the AIC, we selected the four best model, then we decided to choose the one with less coefficients (model $(1,0,1)\times(1,1,1)_s$), as the AIC is known to favors models with too much coefficients. Nevertheless we do not obtain strict white noise for any of those models. As shown in Figure 7, the residuals seem to still have a seasonal component, the Ljung-Box test is rejected and Figure 8 shows that our residuals are not normally distributed.

	AR(1)	MA(1)	SAR(1)	SMA(1)
Coefficients	0.7493	-0.4795	-0.1433	-0.7271
s.e.	0.1642	0.2313	0.0697	0.0810
σ^2 estimated as 10.23	log likelihood:	-1132.72	AIC:	2275.44

Figure 7: Coefficients, log-likelihood and AIC for the SARIMA model $(1,0,1)\times(1,1,1)_{52}$.

In order to try to explain this variation that we supposed due to the variance, we want to try to explain the data on the number of researches by the data of pollens.

3.2 Transfer functions on Allergies with Pollen

We now relate the pollen data to the allergies search counts. We will use transfer functions [3, p. 298] to explain the dependence between the two sets of data. We therefore write a lagged regression model for the output process $\{Y_t\}$, the allergies, as

$$Y_t = \sum_{j=0}^{\infty} \alpha_j X_{t-j} + \eta_t,$$

where $\sum|\alpha_j| < \infty$, and we assume that the input process $\{X_t\}$, the pollen, and the noise process $\{\eta_t\}$ are mutually independent. We will also assume that $\{\eta_t\}$ is stationary.

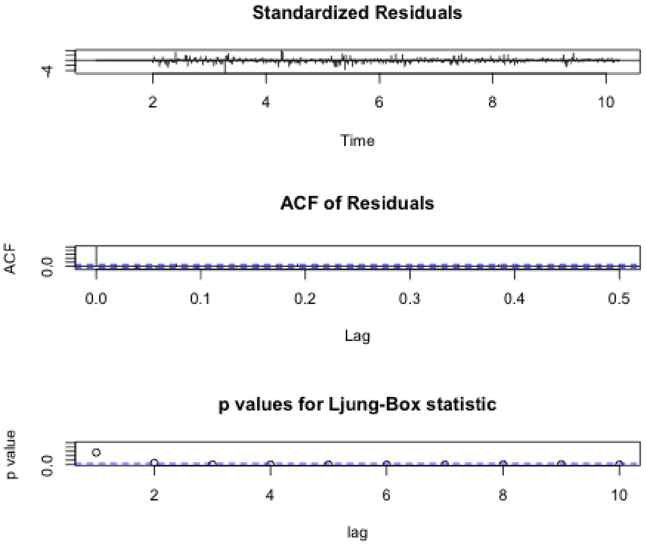

Figure 8: Residuals, their ACF, and p-value of a Portmanteau test for the SARIMA model (tsdiag function in [5]).

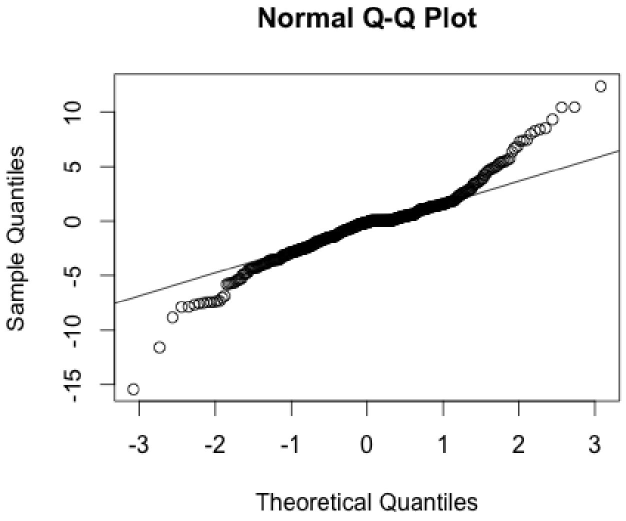

Figure 9: QQplot for the residuals of the SARIMA model).

To obtain a good model that explains the relation between allergies and lagged pollens, we will proceed in 3 steps.

1. We fit a linear model such that allergies are explained by lagged pollens.
2. We extract the residuals and determine wether an additional fit is required. If so, we develop another model to fit these residuals and proceed to step 3.
3. We transform the data $\{Y_t\}$ and $\{X_t\}$ such that we get

$$\tilde{Y}_t = \sum_{j=0}^{\infty} \alpha_j \tilde{X}_{t-j} + \varepsilon_t,$$

where $\{\epsilon_t\}$ is white noise.

3.2.1 Linear Model

The linear model is

$$Y_t = \sum_{j=0}^{9} \alpha_j X_{t-j} + \eta_t, \tag{1}$$

terms after the 9^{th} one were put to zero since they were not significant. By fitting the linear model 1 we obtain residuals on Figure 10. Aside from the first pick on Figure 10, they appear almost stationary with some patterns remaining. Therefore another model needs to be fitted.

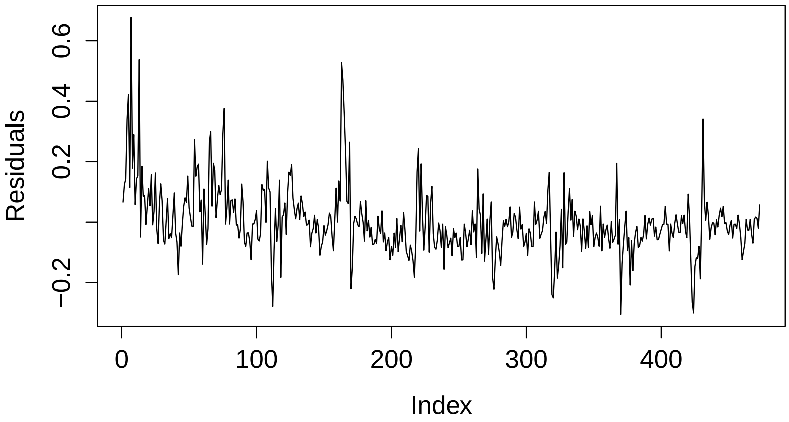

Figure 10: Residuals from the linear model 1.

3.2.2 ARMA on the residuals

We now aim at fitting the residuals seen on Figure 10 with an ARMA model. In fact we will help ourselves with the work done before on the SARIMA modeling of the allergies data. We assume that the seasonal part of our preceding model is explained by the pollen and we take only the ARMA part of it. Consequently, we will fit an AR(1) and MA(1) to the residuals. One can see on Figure 11 that we still have peacks showing seasonality. However, as we want the seasonal part to be explained by the lagged pollen we leave it this way. Further, the signal between two abnormalities of the ARMA residuals seems to behave like white noise.

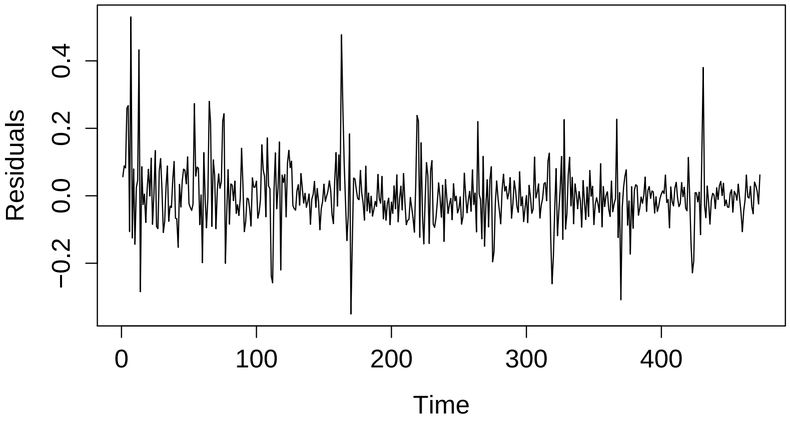

Figure 11: Residuals from the ARMA(1,1) on $\{\eta_t\}$.

3.2.3 Transformed Linear Model

We now want to find $\tilde{Y}_t$ and $\sum_{j=0}^{9} \alpha_j \tilde{X}_{t-j}$. The idea will be to write equation 1 in the following way :

$$\tilde{Y}_t = \frac{\phi(B)}{\theta(B)} Y_t = \frac{\phi(B)}{\theta(B)} \sum_{j=0}^{9} \alpha_j X_{t-j} + \frac{\phi(B)}{\theta(B)} \eta_t = \sum_{j=0}^{9} \alpha_j \tilde{X}_{t-j} + \varepsilon_t$$

With ε_t supposed to be white noise. To transform the data, we must find the coefficients for $\phi(B)/\theta(B)$ in the case of an ARMA(1,1):

$$\frac{\phi(B)}{\theta(B)} = \sum_{j=0}^{\infty} \psi_j B^j = \frac{1 - \phi_1 B}{1 + \theta_1 B}.$$

By solving the equation, we get

$$\begin{aligned}\psi_0 &= 1,\\ \psi_j &= -(\phi_1+\theta_1)(-\theta_1)^{j-1} \quad \forall\, j \geq 1.\end{aligned}$$

After fitting a linear model to the transformed data, we end up with residuals shown in Figure 12.

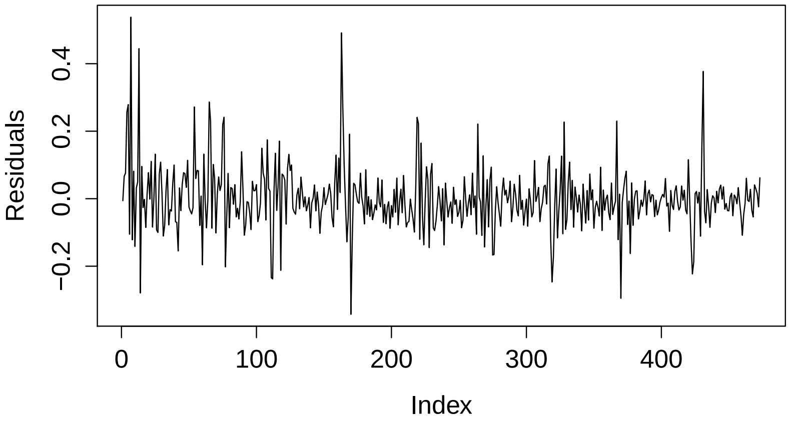

Figure 12: Residuals from the transformed linear model.

It should be noted that due to the applied lag on the pollen and the lack of the data before 2004.01.01, the first 9 rows of the explanatories matrix have been removed. We can observe that Figure 11 and 12 look similar, telling us that the transformation worked. Some seasonality is still observable, as if the pollen did not explain all the seasonal variation. Figure 13 shows the cumulative periodogram for the residuals, and confidence intervals are not crossed.

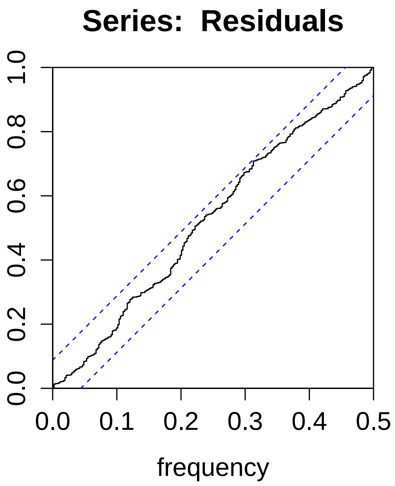

Figure 13: Cumulative periodogram for the residuals of the transformed linear model.

Figure 14 gives tests results on the correlation for the residuals. ACF and portmanteau test support the fact that residuals behave as white noise.

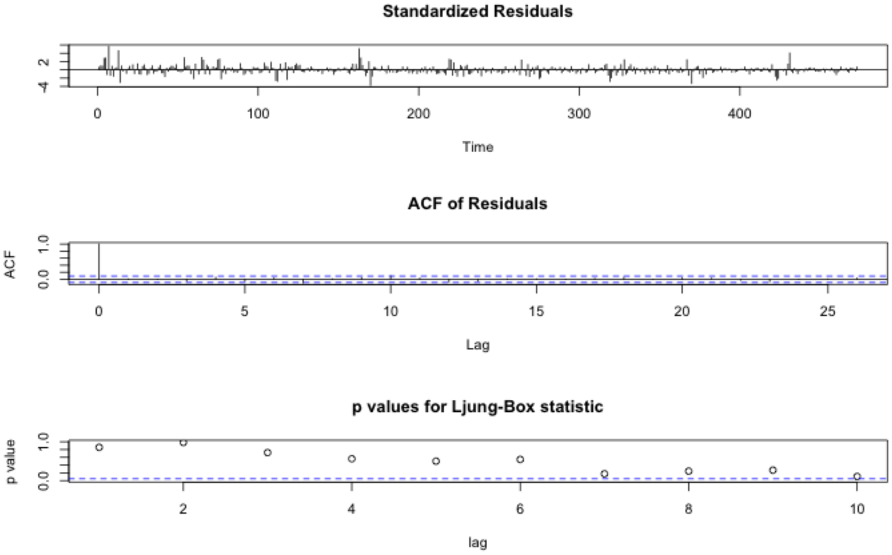

Figure 14: Residuals, their ACF, and p-value of a Portmanteau test for the transformed linear model (tsdiag function in [5]).

Finally, figure 15 shows the coefficients of the transformed linear model. Only the 4^{th} and 5^{th} coefficients are not significantly different than zero at 95% confidence.

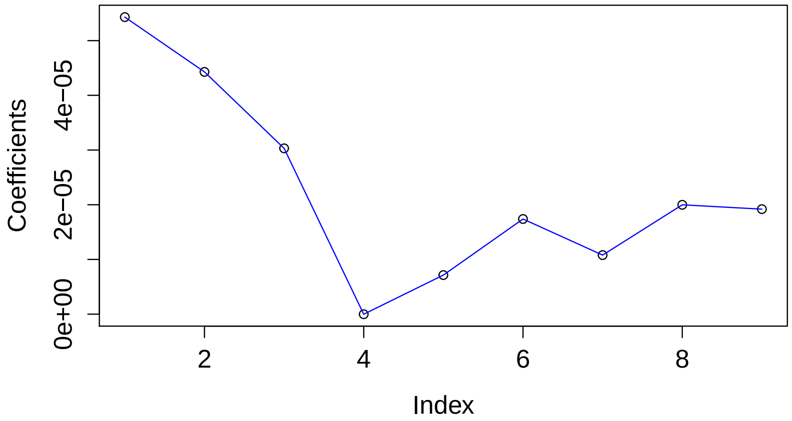

Figure 15: Coefficients for the transformed linear model, the first point is lag zero, the second point is lag one and so on.

4 Discussion

For the first part of our project, we fitted a seasonal ARIMA with parameters $(1, 0, 1) \times (0, 1, 1)_{52}$. According to our method, there was no models with a reasonable number of parameter having a better fit. Can we interpret the coefficients on Table 7 ? We have a positive auto-regressive and a negative moving average part to describe the non-seasonal part. The moving average can simply be interpreted by the way searches counts are given. Indeed, searches counts are gathered weekly, and then averaged on the same period. For the auto-regressive part, it can be interpreted as a simple Markov chain, and would then tell us that Allergies may behave as a Markov chain. On the seasonal part, we have a differentiation and a negative moving average. The differentiation at lag $= 52$ is simply interpreted as the seasonality in the data; however the moving average is hard to explain.

From the tests for the stationarity of the residuals, none of the simple seasonal ARIMAs showed statistically sufficient results. one explanation could be the problematic variance, the picks of allergies in spring do not have the same height (but those differences are also present in the pollen data). Further, the allergy series is very noisy compared to the pollen data. For the later issue, this additional noise in search counts might be due to the fact that the theme allergies is very broad, and queries unrelated to pollen allergies (such as food allergies) will come and inflate the variance of our series.

The second part of our modeling linked pollen counts with searches counts on the topic of Allergies. In the final model, betas where attributed to a lagged version, from zero to nine, of the same pollen data. Figure 15 displays them in order and suggests a smooth decline after lag zero and then bounce back after lag 4. These results match the known pollination period for the trees and graminae studied. Indeed, if trees come first in pollination, graminae come a few weeks later (see appendix 16).

However, we must be careful when we interpret these coefficients. The main motivation for the use of summed pollen counts was that the frequencies of the highest amplitudes are exactly matched between allergies and pollens. But these high peaks in the pollen data, masks the significance of smaller peaks appearing later in the year. This is why the model might twist the information extracted from the data, as it uses more the pollen as reference to the beginning of the spring rather than tree pollen reactions.

We can also comment on the quality of the fit; the first linear model has a R-squared of 0.64 and seven out of ten significant variables (95% confidence). The transformed linear model includes two other parameters (from the ARMA) and has a R-squared of 0.45 and ten out of twelve significant variables (95% confidence). In other words, by including the ARMA modeling, a lagged pollen explanatory becomes significant but we loose in precision according to the R-squared. Another qualitative measure which might be more relevant is the change in the AIC. Indeed, the basic model with non-stationary residuals had an AIC of -2067.8, but with the transformation it is minimized to an AIC of -2220.4.

Another Model could have taken into account differences between types of pollens, and for example having seven explanatory variables for our seven types of pollens. This idea gave us the following results : ash trees, graminae, birch and alder are significantly correlated (99% confidence) to searches counts on Allergies, but hornbeam, hazel and wormwood coefficients are not significant (even at 95% confidence) in the framework of a linear model.

To summarize what this work could bring, we found an important correlation between pollen and searches counts on allergies topic which reinforce the hypothesis that Google Trends does gives a good approximation of the situation on a given topic. We also found some significative lags (such as lags 5 to 8) between the first pick of pollen and allergies.

5 Conclusion

We had in mind to explain allergies search counts. In order to achieve this goal, we fitted two models and explored two different directions. On one hand we used a seasonal ARIMA fitted directly on the search counts. On the other hand a linear model with, as exogenous data, the concentration of pollens. Those two approaches gave us fair results but not as good as we expected.

This can be explain mostly by two reasons. The first reason is that we encountered a major problem with the stabilization of variance. The second reason is that we had not enough information to treat the data on pollens with a pertinent medical interpretation. But even with those problems, the results showed encouraging, and as a sign to pursue the investigation. To develop the research, some ideas would be to find a more efficient way to stabilize the variance and to add, for example, a coefficient to the data on pollen to express the strength of the allergic reactions depending on the type of pollen.

Finally the source of information Google trends, gives a good estimate of the situation on pollen allergies, but we have to keep in mind that it's not directly related to effective cases of these kind of allergies and thus unwanted noise may have affected the results.

References

[1] Federal Office of Meteorology and Climatology MeteoSwiss, *Data from the portal IDAweb for universities*. 7 parameter of pollen in 11 stations from 01.01.2004 00:00 to 30.04.2013 06:50. "http://www.meteosuisse.admin.ch/web/en/services/data_portal/idaweb.html"

[2] Google, *Google Trends*. Limited to the category Health - Health Conditions - Allergies, Switzerland from 04.01.2004 - 10.01.2004 to 17.03.2013-23.03.2013. "http://www.google.com/trends/explore?q=#cat=0-45-419-626&geo=CH&cmpt=q"

[3] A. C. Davidson, *Time series lecture notes*. EPFL, Lausanne, New edition, February 2013.

[4] H. Choi, H. Varian, Predicting the Present with Google Trends, Technical Report, Google Inc.,2011, Available at <http://people.ischool.berkeley.edu/ hal/-Papers/2011/ptp.pdf>

[5] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

A Appendix Google

Explanation of the collection process for Google Trends data as presented by Google [2].

A.1 How is the data derived?

Google Trends analyzes a portion of Google web searches to compute how many searches have been done for the terms you've entered, relative to the total number of searches done on Google over time. This analysis indicates the likelihood of a random user to search for a particular search term from a certain location at a certain time. Keep in mind that Trends designates a certain threshold of traffic for search terms, so that those with low volume won't appear. Our system also eliminates repeated queries from a single user over a short period of time, so that the level of interest isn't artificially impacted by these type of queries. Say you've entered the search term tea, setting your location parameter to Scotland, and your time parameter to March 2007. In order to calculate the popularity of this term among users in Scotland in March of 2007, Trends examines a percentage of all searches for tea within the same time and location parameters. The results are then shown on a graph, plotted on a scale from 0 to 100. The same information is also displayed graphically by the geographic heat map.

A.2 Is the data normalized?

Yes. All the results in Google Trends are normalized, which means that we've divided the sets of data by a common variable to cancel out the variable's effect on the data. Doing so allows the underlying characteristics of the data sets to be compared. If we didn't normalize the results and displayed the absolute rankings instead, data from regions generating the most search volume would always be ranked high. Let's consider the following examples to highlight some of the key points of normalization: *Canada and Fiji show the same percentages for the term 'hotel.' Does this mean that they have the same amount of search volume for that term? Just because two regions show the same percentage for a particular term doesn't mean that their absolute search volumes are the same. Data from these two regions - with significant differences in search volumes - can be compared equally because the data has been normalized by the total traffic from each respective region. So, we can assume that users in both Fiji and Canada are equally likely to search for the term hotel. *New York doesn't appear on the list for the term 'haircut.' Does this mean that people in New York don't search for this term at all? Remember, Google Trends shows the likelihood of users in a particular area to search for a term on Google on a relative basis. So, just because New York isn't on the top regions list for haircut doesn't necessarily mean that people there don't search for that term at all. Consider the following scenarios. It could be that people in New York: don't use Google to find a barber or hair salon use a different term for haircut-related searches search for so many other topics unrelated to haircuts, that searches for haircut comprise a small portion of the search volume from New York as compared to other regions

A.3 How is the data scaled?

The data is displayed on a scale of 0 to 100. To arrive at those values, we first normalize the data. After normalization, we divide each point on the graph by the highest value and then multiply by 100. Different plots are not comparable unless they share the same original highest value before scaling. For example, let's suppose that interest for the term skiing surged in the month of November in Sweden. Our system designates that peak as 100. Now let's suppose that interest decreased significantly in December, where the next highest peak was approximately half of what it was in November. That peak would then be designated as 50, and so on.

A.4 What do the numbers on the graph mean?

The numbers on the graph reflect how many searches have been done for a particular term, relative to the total number of searches done on Google over time. They don't represent absolute search volume numbers, because the data is normalized and presented on a scale from $0 - 100$. Each point on the graph is divided by the highest point, or 100. When we don't have enough data, 0 is shown. Read more about how we scale and normalize the data. When you apply the Category filter, you'll see another graph. This graph will show the change over time as a percentage of growth, with respect to the first date on the graph (or the first date that has data). That's why you'll notice that instead of a 0-100 label on the y-axis of the category comparison graph, you'll see a label with a range of -100% or $+100\%$, and a starting point of 0.

B Appendix Pollens

Explanation of the collection process for pollen data as presented by Meteo Suisse [1].

B.1 How pollens are captured and measured?

Each of the 14 measurement stations is equipped with a sensor volumetric pollen. This pollen trap sucks using a pump ten liters of air per minute through an opening of 14 x 2 mm. Behind the suction slot rotates a drum covered with a strip of cellophane coated with silicone. Pollen and other organic and inorganic particles contained in the air are adhered to the tape. The drum is replaced every week and sent to one of the two analyzes located in Zurich and Payerne centers. This is where the band is divided into daily preparations. Pollens are then identified and counted under a microscope. On the support, there is, in addition to other organic particles such as mold spores, inorganic particles such as Saharan dust or soot. With this long and careful manual analysis, all data pollen a week 14 measurement stations are available from the following Wednesday and available online for you in the Swiss pollen newsletter.

B.2 Data of pollens

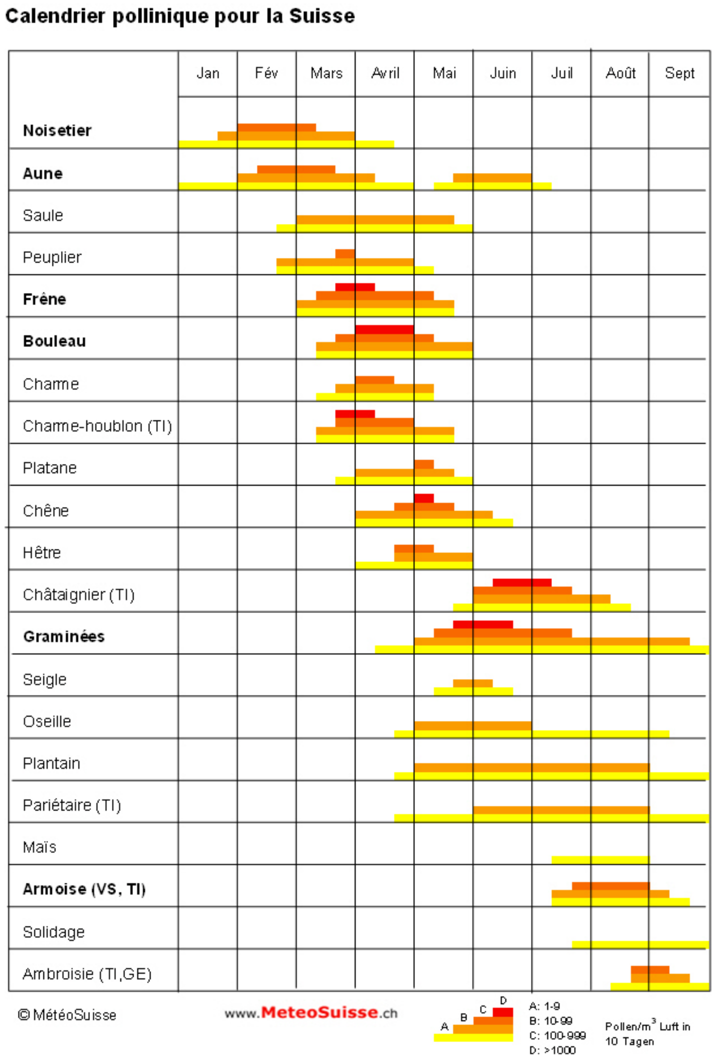

Figure 16: Annual chart of pollens.

Pollen code:

- X1343 noisetier [No]
- X1331 charme [No]
- X1352 frene [No]

- X1428 armoise [No]
- X1469 graminees [No]
- X1323 bouleau [No]
- X1315 aune [No]

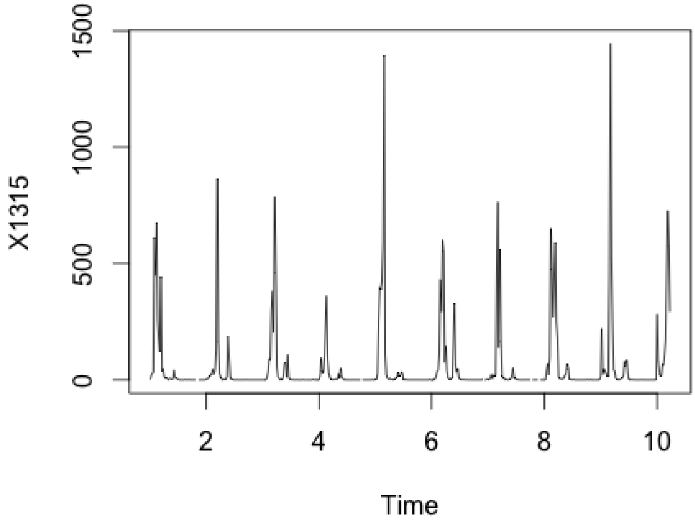

Figure 17: Average of pollen parameter 1315 (Alder) on stations.

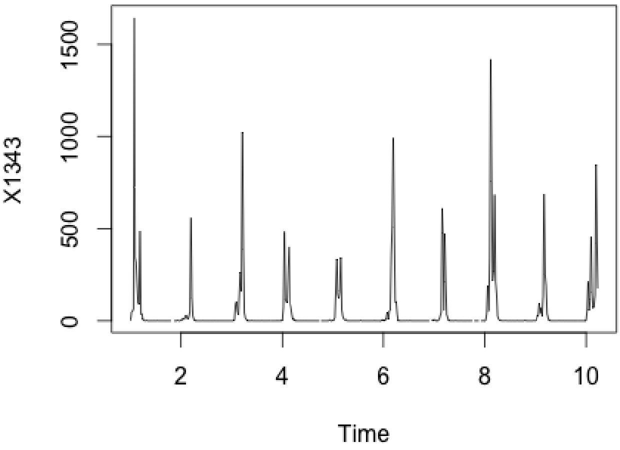

Figure 18: Average of pollen parameter 1343 on stations.

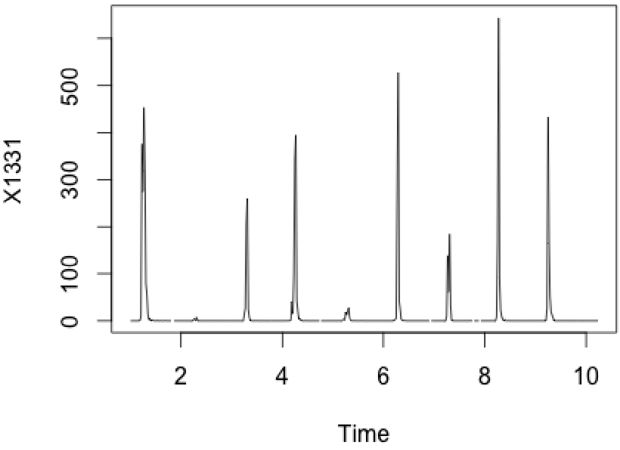

Figure 19: Average of pollen parameter 1331 on stations.

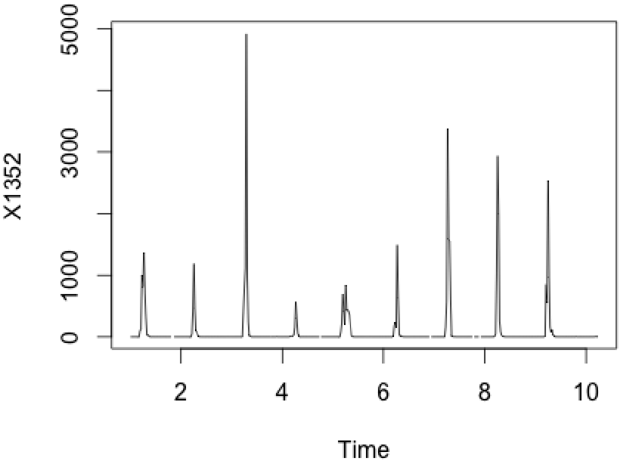

Figure 20: Average of pollen parameter 1352 on stations.

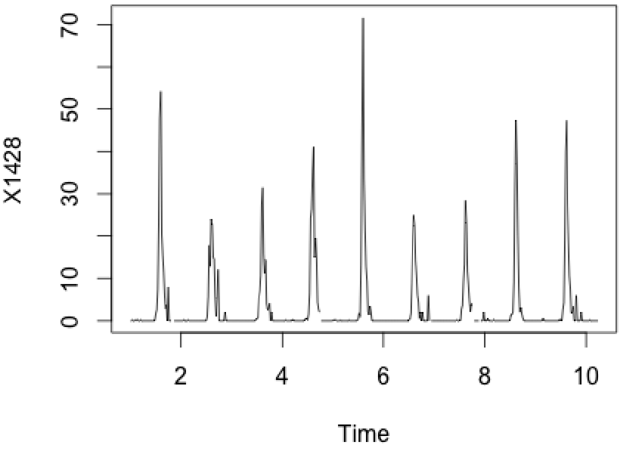

Figure 21: Average of pollen parameter 1428 on stations

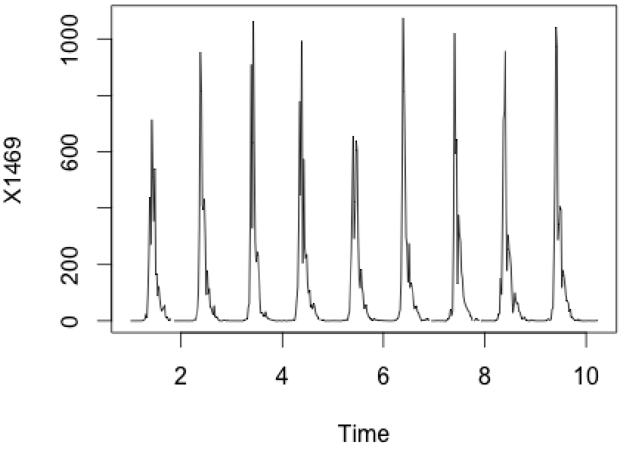

Figure 22: Average of pollen parameter 1469 on stations.

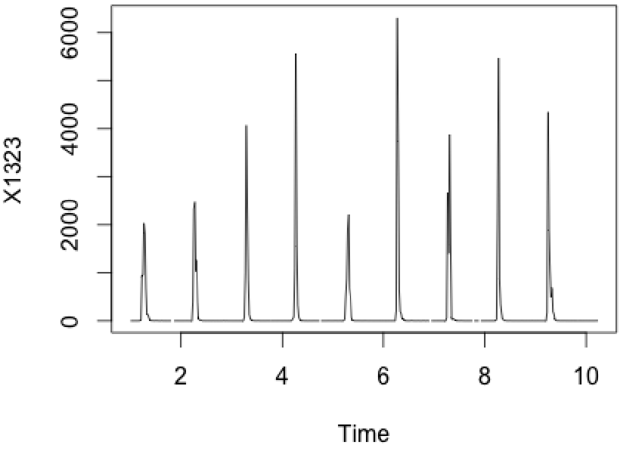

Figure 23: Average of pollen parameter 1323 on stations.

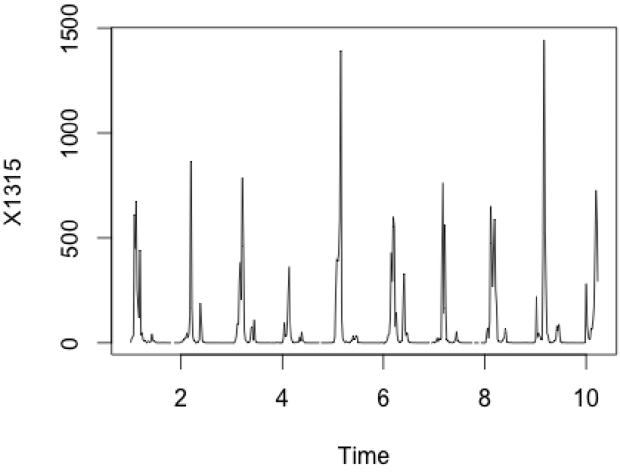

Figure 24: Average of pollen parameter 1315 on stations.