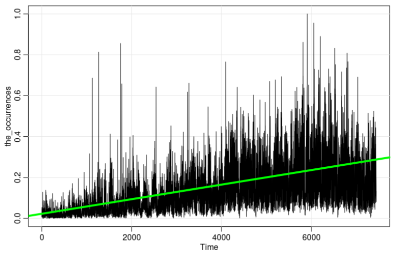
1 Introduction

This project analysis a large set of data from Internet Engineering Task Force (IETF) e-mailing lists, ranging from November 1997 until May 2018. From each of the e-mailing lists, each of the e-mails were extracted and searched for keywords *the*, *security* and *privacy*, to see how the organisation increases or decreases its focus on these topics over time, and whether such focus shifting can be modelled with any of the time series models learned in this course. The Python-tools used to download and extract data from the e-mails can be found on Datactives Bigbang Github depository.[1]

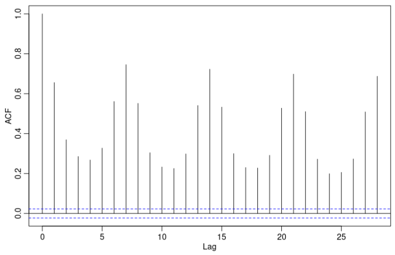
The keyword *the* is used to normalise the range of occurrences of other keywords to the set $[0, 1]$. Because the determinate article will occur in most sentences in the English language, the *the*-scale also gives an indication of the total amount of words communicated on the e-mailing lists.

In Fig. 1 I have presented occurrences of the word *the* in e-mails over the given time-period, and linearly regressed *the*-occurrences with respect to time. I've also extracted an auto-correlation function for the time-series (Fig. 2). The e-mailing lists appear to be more used over time, and exhibit seven-day cycles. Indeed, the seven-day cycle can be verified by running the FFT-transformation on the *the*-series and studying the periodogram (see Fig. 3).

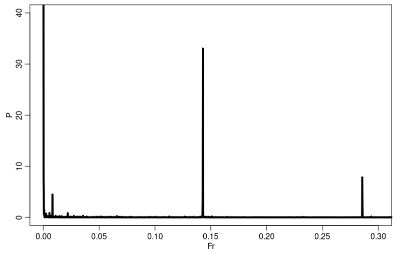
Mathematical notation is introduced in this document to more easily describe transformations (and remember which particular transformations have been performed on the series at any given stage of the report). However, $\mathcal{D}_F \tilde{z}_s$ and *normalised (with respect to "the"), filtered, differenced security-series* might be used interchangeably, for example.



Figur 1: Number of *the*-occurrences, and a green line indicating an upwards trend.



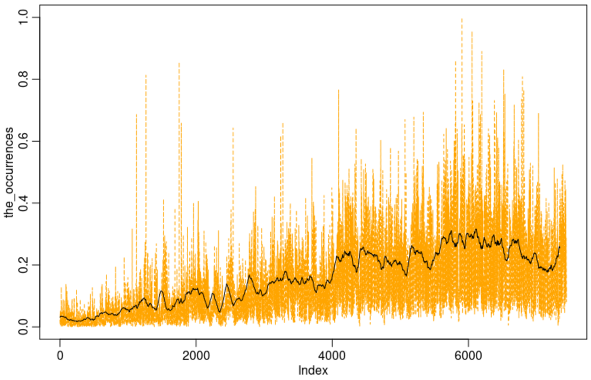
Figur 2: The ACF indicates seven-day cycles.



Figur 3: The periodogram shows the seven-day cycle, and two resonance cycles.

[1] See `https://github.com/datactive/bigbang`

2 Filtering the time series



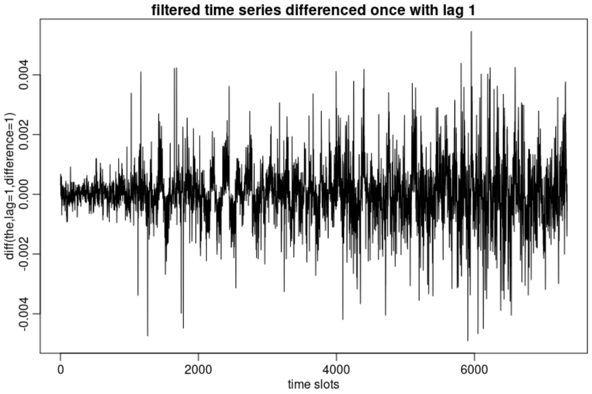
Figur 4: Filtered (black) and original (orange) series.

I want to start with filtering out the seven-day cycles. This requires a low-pass filter, as I'm hoping to preserve longer trends in the series. I use the modified Daniell kernel with filter weights that are multiples of 7, similar to what is described in the book. After testing different multiples, I settle for using filter weight 49. The results can be seen in Fig. 4, where the original series is displayed in orange, and the filtered series is shown on top as a black line.

The impression of a linearly increasing trend persists. To see if the remainder of the time series is randomly distributed, I tried differencing. Letting the data in the filtered *the*-series be $t_1, t_2, ...$ and expressing time with s, I let $x_s = t_s - t_{s-1}$ and seek to perform stationarity tests on $\{x_s\}$. Exploratory analysis of $\mathcal{F}[x_s]$ in Fig. 5 looks promising.

Indeed, the KPSS test and the ADF test both indicate that the series is stationary. From the KPSS test, I get a KPSS level of 0.033 and a p-value of 0.1 for a level test. This is not sufficient to reject the null hypothesis that the series is stationary. From the ADF test, I have a p-value of 0.01, and I should reject the hypothesis that the series is not stationary.

Arguably, these results are not very interesting. The IETF standardises protocols and features for a popular technology (the Internet), the deployment of which is growing around the world. It is expected that the IETF e-mailing lists traffic would grow over time, just as it made sense that there were seven-day cycles in an organisation dominated by professionals (who work during the week, but perhaps not during weekends).



Figur 5: Differenced with lag 1.

The techniques used above can, however, be re-used on other word-counts extracted from the same e-mailing lists. In the introduction it was mentioned that use of the words *security* and *privacy* were investigated, and that the *the*-series was collected only as a normalising series. For these words it possible to have several suspicions: as media focus on security and privacy goes up over time, we might expect to see a bigger focus on security and privacy in IETF communications (with some lag). We could also suspect to see such increased focus following new privacy or security oriented laws. Perhaps we suspect that privacy discussions do not exhibit as strong tendencies of periodicity, because privacy is more often cared for by idealistic people than is security.

3 *Privacy* and *Security* mentions

Let $\{p_s\}$ be the mentions of privacy at each time slot s, $\{z_s\}$ be the mentions of security and $\{t_s\}$ be the mentions of the word *the*. For each s we map $\{p_s\}$ on to $[0, 1]$ using the transformation

$$\tilde{p}_s = \frac{p_s}{\max\{t_s\}}$$

and similarly for $\tilde{z}_s$ and $\tilde{t}_s$. The time $s \in [1, 7442]$ for this data set, or begins in 1997-11-01 and ends approximately on 2018-05-04 (note that neither the control word *the* nor any of the other words were recorded on some days).

We will be studying two time-series and their transformations:

1. $\tilde{z}_s$, the series of mentions of the word *security* normalised on the scale of how many times the word *the* was mentioned each day.

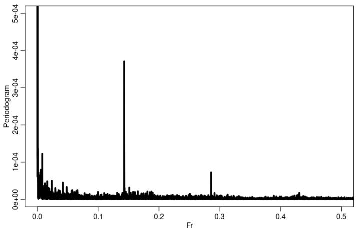
2. $\tilde{p}_s$, the series of mentions of the word *privacy* normalised on the scale of how many times the word *the* was mentioned each day.

The notation $\tilde{t}_s$ means simply the time-series for *the* normalised with respect to itself. We will also use $\mathcal{F}$ to denote the low-pass filter consisting of a modified Daniell kernel with parameters 7, 7, and eventually $\mathcal{D}_F$ to denote the application of both the low-pass filter and differencing with lag 1.

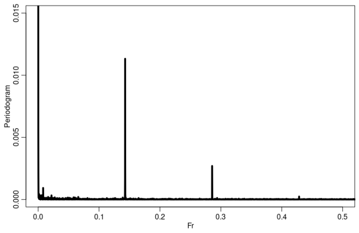
3.1 Periodicity prejudices confirmed

It turns out that while both $\tilde{z}_s$ and $\tilde{p}_s$ exhibit seven-day cycles, the periodogram is much more clean for $\tilde{z}_s$ (cf. Fig. 7). This conforms with the expectation that privacy advocates are more happy about working in the weekend (note the clutter in Fig. 6). Also *security* is a more often used word than *privacy* in IETF discussions, a fact observed by plotting the $\tilde{z}_s - \tilde{p}_s$ series too. This is visible in an unfiltered plot too, but the filtered plot in Fig. 8 makes it easy to see. Such a trend is not unreasonable, given that security as such has broader applications than privacy.

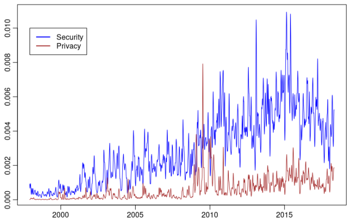
In Fig. 8 there is an interesting spike in mentions of privacy around years 2009–2010. During this period *privacy* is the more frequently occurring word. There is another spike around 2014–2015, although in this period *privacy* is not more mentioned than *security* as there is a spike also in mentions of *security*. Notably, after the 2009 spike



Figur 6: Exhibit A: the *privacy*-series periodogram. `xlim=c(0,.5),ylim=c(0,5e-04)`



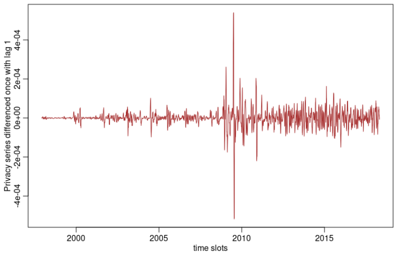
Figur 7: Exhibit B: the *security*-series periodogram. `xlim=c(0,.5),ylim=c(0,.015)`



Figur 8: *Security* is more frequently mentioned word than *privacy*. Timescale is 1997–2018.

privacy appears to have become a consistently more present topic. However, if we were to plot $\tilde{t}_s - \tilde{z}_s$ and $\tilde{t}_s - \tilde{z}_s$, we would also see that the total amount of communications (represented by $\tilde{t}_s$) increases faster than $\tilde{z}_s$ or $\tilde{p}_s$.

Since $\mathcal{F}[\tilde{z}_s]$ and $\mathcal{F}[\tilde{p}_s]$ both appear to exhibit a linear trend, we can also try differencing them to get a stationary series. By repeating the same plotting as in Fig. 5, both filtered, differenced series appear to exhibit greater variance over time. The results can be seen in Fig. 9 and Fig. 10.

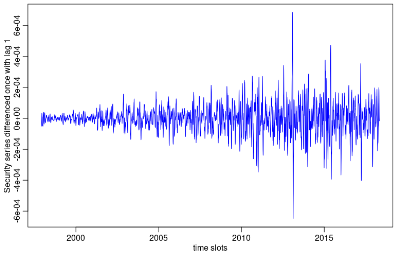


Figur 9: $\mathcal{D}_F[\tilde{z}_s]$. Time-scale is 1997–2018.

3.2 Developments in the noise

In this section, we will refer to the filtered, differenced time series as $\mathcal{D}_F[\tilde{z}_s]$ and $\mathcal{D}_F[\tilde{p}_s]$. They can be seen in Fig. 9 and Fig. 10, respectively.

Now that we have filtered, differences time series for both *security* mentions and *privacy* mentions, we can start looking at differences in the approach to these two words as used on IETF e-mailing lists. KPSS and ADF tests both confirm we are looking at series that are somehow stationary (KPSS indicates we cannot reject the hypothesis of stationarity, and ADF indicated we should reject the hypothesis of non-stationarity).



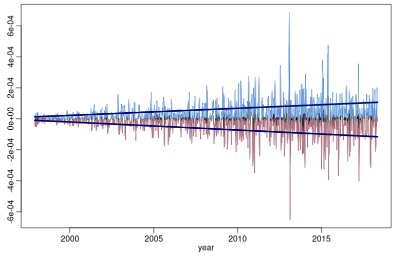
Figur 10: $\mathcal{D}_F[\tilde{p}_s]$. Time-scale is 1997–2018.

3.2.1 $\mathcal{D}_F[\tilde{z}_s]$: mentions of security

The 95% t-confidence interval for the mean μ contains 0, so we can say the time-series oscillates around 0. The variance is not constant in the series, but appears to be growing over time.

ACF and PACF are trailing off slowly for $\mathcal{D}_F[\tilde{z}_s]$ (not displayed). By separating negative and positive values in the series, we can try investigating whether negative values are increasing at a slower or a faster rate than positive values. When we fit the positive values of $\mathcal{D}_F[\tilde{z}_s]$ to a linear model, we get $z_s^+ = 0.000013 + 0.000000013s$. The negative fit is $z_s^- = -0.00001 - 0.000000014s$. The results can be seen in Fig. 11.

Residual analysis for both fits indicates that neither fit is particularly good (rather than being normally distributed, both the residuals and the standard residuals are linearly distributed on a negative slope).



Figur 11: Linear regressions on Var[$\mathcal{D}_F[\tilde{p}_s]$]. Time-scale is 1997–2018.

It is perhaps natural to consider whether an ARMA+GARCH model could be used to describe the data. To test this in R, we scale up the *security*-series in the following way: $100 \times \mathcal{D}_F[\tilde{z}_s] = \mathcal{D}_F[\tilde{z}_s]_{100}$. Otherwise, the values become too small for the `garchFit`-function to terminate.

Running `garchFit(∼arma(1,0)+garch(1,0), data=diffseclow_100)`, where `diffseclow_100`$= \mathcal{D}_F[\tilde{z}_s]_{100}$, I found the parameters $\hat{\theta}_0 = 0.97$,$\hat{\alpha}_0 = 6e - 7$ and $\hat{\alpha}_1 = 1.0$. The residuals and standard residuals appear to be distributed almost the same as $\mathcal{D}_F[\tilde{z}_s]$, rather than normally, so that does not inspire confidence in the model. The parameter $\hat{\alpha}_1 = 1.0$ makes it impossible to use `garchFit` in R.

Over many attempted ARMA+GARCH or GARCH models to describe the *security*-series, generally the $\hat{\alpha}_1$-parameter ends up fairly large, while β-parameters end up small. AR and MA-parameters end up approaching 1, either both together or one at a time. We may suspect the volatility of variance is volatile, rather than linearly increasing over time.

The development of the variance may in fact not be stochastic at all, but rather a mostly deterministic function (except for perhaps some noise) of the amount of communications passing through the e-mailing lists at any given day. The assumption of independence of observations may not hold: if a discussion is launched on either privacy or security on a given day, that discussion may continue for several days between e-mailing list participants causing the number of mentions on several consecutive days to be dependent on each other.

3.2.2 $\mathcal{D}_F[\tilde{p}_s]$: mentions of privacy

In the analysis of $\mathcal{D}_F[\tilde{p}_s]$ it might be useful to consider the series in two separate intervals. In Fig. 10 we see a clear shift in behaviour around 2009–2010, before which the behaviour is different from the behaviour exhibited after this event.

We first select the first 3900 observations in $\mathcal{D}_F[\tilde{p}_s]$, and see whether it conforms to any particular model. The entire series needs to be scaled up by 1000, to ensure that the R-function `garchFit` converges. Because both ACF and PACF are trailing off, we proceed with searching for ARMA(a,b)+GARCH(c,d) models, where $a, b, c, d \in \{0, 1\}$. After some analysis, we find an ARMA(1,1)+GARCH(1,1) model, for which residual analysis looks relatively okay. The output from R is:

```
Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
mu     2.496e-05   1.815e-05    1.375    0.169
ar1    9.501e-01   5.014e-03  189.500  < 2e-16 ***
ma1    9.938e-01   1.251e-03  794.081  < 2e-16 ***
omega  2.229e-09   5.675e-10    3.928 8.56e-05 ***
alpha1 1.783e-01   8.817e-03   20.218  < 2e-16 ***
beta1  8.677e-01   4.311e-03  201.299  < 2e-16 ***
```

Residuals maintain a more normal-like distribution which a much lower rate of auto-correlation than the original time-series. The ACF for the squared standardised residuals is mostly cut off at 1 (with a small spike at 14). While information criteria are a bit higher than for simpler models, visual inspection of the residuals are hopeful. Residual tests, however, come far from indicating normally distributed residuals.

Because we found a high value for $\hat{\beta}_1$ and a lower value for $\hat{\alpha}_1$, with the AR and MA-parameters being around $0.95 - 0.99$, the variances changes less unexpectedly for this series (i.e. it is less volatile).

The behaviour of the ACF and PACF for the segment of $\mathcal{D}_F[\tilde{p}_s]$ that runs from the middle of 2010 until now (approximately equalling observations at 4250 to 7413) is similar to that of the pre-2009 series. They both slowly trail off. After testing a few different models, and performing residual analysis through using the `plot`-function in R, it appears that a simple GARCH(1,1)-model may be the best model for the data.

```
Error Analysis:
         Estimate  Std. Error  t value Pr(>|t|)
mu     -2.165e-03   6.453e-04   -3.355 0.000794 ***
omega   3.793e-05   5.185e-06    7.316 2.56e-13 ***
alpha1  9.361e-01   3.354e-02   27.908  < 2e-16 ***
beta1   1.280e-01   1.684e-02    7.597 3.04e-14 ***
```

The information criteria (AIC, BIC, SIC) are all around -4.3, which is lower than for the other models mentioned above. The standard residual tests are not encouraging.

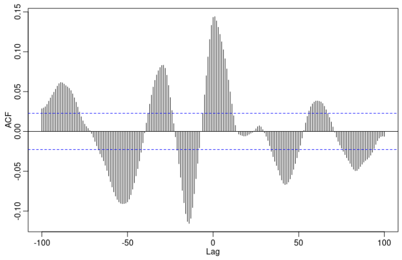
```
Standardised Residuals Tests:
                                Statistic p-Value
 Jarque-Bera Test   R    Chi^2  89.60451  0
 Shapiro-Wilk Test  R    W      0.9306203 0
 Ljung-Box Test     R    Q(10)  6907.455  0
 Ljung-Box Test     R    Q(15)  7892.148  0
```

In this model too, the variance is volatile. The tendency for a big $\hat{\alpha}_1$ and a small $\hat{\beta}_1$ persists over various attempts to fit ARMA(a,b)+GARCH(c,d) models to the data.

Note that both segments of $\mathcal{D}_F[\tilde{p}_s]$ were scaled up 1000 times prior to running the `garchFit` function, in order to avoid problems of terminating the fitting algorithm for very small numbers.

3.2.3 $\mathcal{D}_F[\tilde{p}_s]$ and $\mathcal{D}_F[\tilde{z}_s]$ together

In the book, we are given tools to model the cross-covariance of two time series. The cross-covariance of $\mathcal{D}_F[\tilde{p}_s]$ and $\mathcal{D}_F[\tilde{z}_s]$ is exhibited in Fig. 12. It appears as if discussions on privacy are at least partially leading discussions on security, while the opposite is true to a smaller extent.



Figur 12: Cross-correlation between $\mathcal{D}_F[\tilde{p}_s]$,$\mathcal{D}_F[\tilde{z}_s]$.

4 Conclusions

The z_s series is not easy to describe with the models I have tried above. While work-week cycles and long-term positive trends can be filtered out, the remaining noise appears not to lend itself to description with the methods I have tried.

For the first segment of the p_s series I was able to find an ARMA(1,1)+GARCH(1,1) model that resembled the normalised, filtered, differenced time series. For this segment, variance was also found to be generally less volatile. For the second segment of the p_s series, I was able to fit a GARCH(1,1) model and able to confirm a more volatile variance over many different attempted model fits.

In all these cases, residual tests indicated that the residuals were less than normally distributed. However, analysing the residuals for the chosen models also showed that squared standardised residuals were as good as uncorrelated between lags, while standardised residuals had slightly more auto-covariance but significantly less than the series themselves.

The cross-correlation indicated dependencies between the number of mentions of privacy and the number of mentions of security. In particular, mentions of privacy lead the mentions of security.

5 References

1. Datactive, Bigbang. `https://github.com/datactive/bigbang`

2. Shumway, R. H., & Stoffer, D. S. (2017). Time series analysis and its applications: With R examples (4th ed.). New York: Springer.