Examination of a Demand Driven Real Estate Index Based on Search Query Data

Jasmin Maag, 10-714-715

12 October 2018

Abstract. An up-to-date index that is able to reflect the demand for real estate is needed not only in practice, but as well for theoretical applications. Based on Google search inquiries, indices for the search term "Wohnung" in the cantons Zurich and Bern are constructed. These indices are then approached with the classical decomposition model.

Introduction

Real estate is an indispensable good of an economy. Besides serving as a necessary living space for private individuals, construction and maintenance generate a large number of jobs. And as a financial asset—either as a private investment or grouped into real estate funds—it can create significant added value for the investor, and at the same time have the power to throw an entire economy into a crisis: real estate-bubble warnings are a recurring subject in the media all around the world. However, despite the importance of the real estate market, there is not a lot of information about the interplay of its supply and demand. For example, new investment projects are often justified by assessing demand for real estate based on retrospective measures such as past market transactions. In the American real estate market, researchers have therefore begun to include search query data in order to improve the quality of pricing models. For example, Wu and Brynjolfsson (2015) show that the number of Google search inquiries for "real estate agencies" and "real estate listings" have a significant influence on the number of home sales and on house prices.
In this short paper two demand-driven real estate indices for the cantons Zurich (ZH) and Bern (BE) are constructed based on Google search inquiries for the term "Wohnung". In a second step the indices are modeled with a classical decomposition approach.

1 Data

The website Google Trends allows users to analyze the popularity of specific search terms in a country and time period. The daily number of search inquiries for a time period is normalized to a percentage of the highest number in this time period. That is to say, the number of inquiries of the day with the highest number of inquiries in a time period is set to 100 and the number of the other days relative to that (Google Trends (2018)). In this paper the daily number of search inquiries for the term "Wohnung" is used as a data basis. For this 1,098 daily files were downloaded, covering the time period 31 August 2015 to 31 August 2018. Since the daily number showed a lot of noise, the daily data was smoothed by taking a moving average of the previous 7 days. Hence, the final indices for the cantons ZH and BE cover the period 6 September 2015 to 31 August 2018. An overview of the

two indices is shown in Figure 1. As can be seen, both indices show a slightly increasing trend over the three years under observation.

2 The Classical Decomposition Model

The classical decomposition model is based on the idea that any non-stationary time series can be separated into non-stationary and stationary components. In this paper, an underlying additive model is assumed. Hence, the model is

$$X_t = T_t + C_t + S_t + E_t, \tag{1}$$

where T_t describes the trend, C_t the cyclical component, S_t the season, and E_t the error component (Brockwell and Davis (1991)). Regarding the two index series for ZH and BE, for T_t a linear trend model is chosen, for S_t a simple seasonal average, and E_t is modeled as an ARMA(p, q) process. Since the data covers only three years, C_t was assumed to be not present. The decomposition model was calculated for the period 6 September 2015 to 24 August 2018. The period 25 to 31 August 2018 was left out in order to forecast future values with the final models.

3 Application

In a first attempt a linear trend model is chosen for T_t. In practical applications it is common to use a polynomial function up to order two (Schlittgen (2001)). Hence, three different polynomial functions m_t are fitted to the indices,

1. $m_t = \beta_1 + \beta_2 t$
2. $m_t = \beta_1 + \beta_3 t^2$
3. $m_t = \beta_1 + \beta_2 t + \beta_3 t^2$

The models are fitted with a linear regression and the parameters are estimated with OLS. In line with Schlittgen (2001) the model with the lowest BIC is chosen to represent T_t. For ZH the trend model $x_t = \beta_1 + \beta_2 t + \beta_3 t^2$ showed the lowest BIC with 5915.25, and for BE the model $m_t = \beta_1 + \beta_3 t^2$ with a BIC of 6124.47. An overview of the modeled trend functions is given in Table 1. The parameters are significant on a 5 percent significance level. Further, a plot of the trend function alongside the index values is provided in Figure 2.

In order to model the season, the modeled trend values have to be subtracted from the original data. Figure 3 shows a yearly overview of the de-trended indices ZH and BE. Although the de-trended series show no obvious seasonality, S_t was calculated as a simple seasonal average in order to further reduce the noise in the data. A graphical overview of the de-trended and de-seasonalized indices is given in Figure 4. As can be seen, the data no longer shows any significant deviation from stationarity, and the error term is ready to be modeled.

As mentioned in the beginning, E_t is approached with an ARMA(p, q) model. To find the best combination of the orders p and q, the dependency structure of the error series is assessed by plotting the autocorrelation function (ACF), partial autocorrelation function (PACF), and by calculating the vector-correlations. Mathematically the vector-correlation for orders p and q is defined as

$$\lambda(p, q) = (-1)^p \cdot \frac{\kappa_{p+1,q+1}}{\kappa_{p+1,0}}, \tag{2}$$

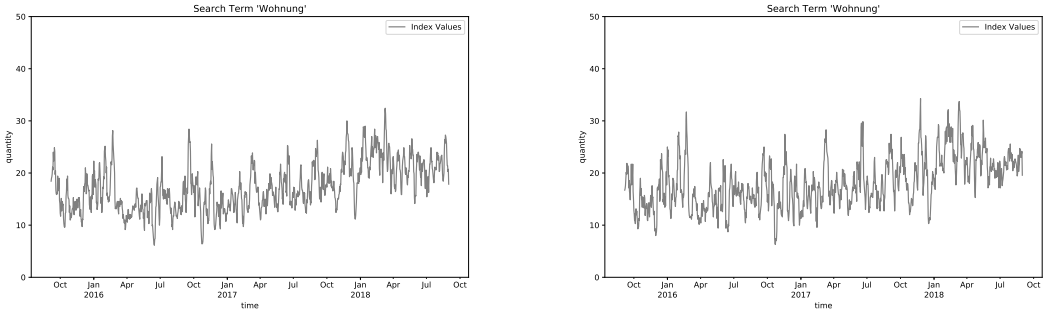

Figure 1: Plots of the indices ZH (left) and BE (right). The x-axis describes the time from 6 September 2015 to 31 August 2018, and the y-axis the daily popularity of the search term "Wohnung". Source: own elaboration based on Google Trends (2018).

	coeff	pvals	conf_lower	conf_higher
const	15.791039	4.293426e-264	15.132997	16.449081
time	-0.007115	7.243151e-07	-0.009916	-0.004314
time^2	0.000014	1.101933e-25	0.000011	0.000016

	coeff	pvals	conf_lower	conf_higher
const	15.222830	0.000000e+00	14.860039	15.585620
time^2	0.000007	2.006833e-83	0.000007	0.000008

Table 1: Overview of the fitted polynomial trend functions for the indices ZH (top) and BE (bottom). Source: own elaboration based on Google Trends (2018).

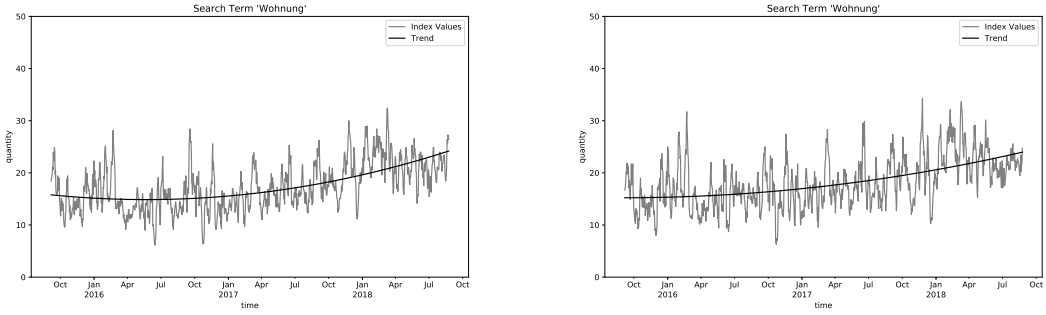

Figure 2: Plots of the indices ZH (left) and BE (right) with their polynomial trend functions. The x-axis describes the time from 6 September 2015 to 24 August 2018, and the y-axis the daily popularity of the search term "Wohnung". Source: own elaboration based on Google Trends (2018).

where $\kappa_{p,q}$ describes the determinant of a Toeplitz matrix with the ACF value $\rho(q)$ as main diagonal element (Schlittgen and Streitberg (2001)). The vector correlations should take on values equal or close to 0 for combinations of p and q, where p and q are equal to or higher than the real orders p and q of the underlying ARMA(p, q) process (Schlittgen and Streitberg (2001)). The vector-correlations for combinations of $p \leq 6$ and $q \leq 6$ are given in Table 2. As can be seen, both vector-correlation tables show a "zero" block starting at $p = 3$ and $q = 2$. This would suggest a p and q combination of 3 and 2 for both indices. However, the ACF and PACF plots in Figures 5 and 6 show significant values up to lag 6, and 14 respectively. Thus, to find the best combination of p and q, the corrected Akaike information criterion ($AICc$) is calculated. The $AICc$ is defined as

$$AICc = AIC + \frac{2(p+q)(p+q+1)}{n-p-q-1}, \tag{3}$$

where AIC is the normal Akaike information criterion and n the number of observations. An overview of the $AICc$ values for combinations of $p \leq 7$ and $q \leq 7$ is given in Table 3. For both indices the lowest values are found for $p = 1$ and $q = 6$. Hence, for ZH and BE an ARMA$(1, 6)$ is modeled to the error series. An overview of the found ARMA$(1, 6)$ is given in Table 4. Based on a 5 percent significance level, both error models show significant parameters.
In order to asses the error model's fit, the residuals of the error model should be inspected for randomness, and if they show any remaining dependency structure. For this, the ACF and PACF values of the residuals are plotted in the Figures 7 and 8. Neither the ACF nor the PACF plots overstep the 95 percent boundaries. In conclusion, the fitted ARMA$(1, 6)$ model for the error component of the indices ZH and BE provides a good fit.

In a final step the overall fit of the calculated decomposition model is assessed. To do so, the mean absolute percentage error ($MAPE$) for the in- and out-of-sample period was calculated. The general $MAPE$ formula is

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{x_t - \hat{x}_t}{x_t} \right|, \tag{4}$$

where x_t are the observed values, and $\hat{x}_t$ the model's estimated values for a time period of length n. For both indices the in-sample $MAPE$s are fairly low with approximately 0.006. The out-of sample $MAPE$ is 1.24 for ZH and 0.69 for BE. Additionally, the original values are plotted alongside the modeled values in Figure 9. The modeled values seem to fit pretty closely to the original values.

Conclusion

The task of this paper was to construct an up-to-date index for the demand for real estate. Based on Google search inquiries for the term "Wohnung" two indices for the cantons Zurich and Bern were constructed and modeled with a decomposition approach. Overall, the modeled components of the decomposition model showed fairly good results with parameter p-values far below a 5 percent significance level. However, it is questionable if it was indeed necessary to calculate a seasonal component, since a season did not show to be prominently present in the yearly overview of the de-trended indices. Further, the close fit of the modeled values to the original values suggests that overfitting might be present. Improving the modeled indices and investigating the problem of overfitting will be left to future research.

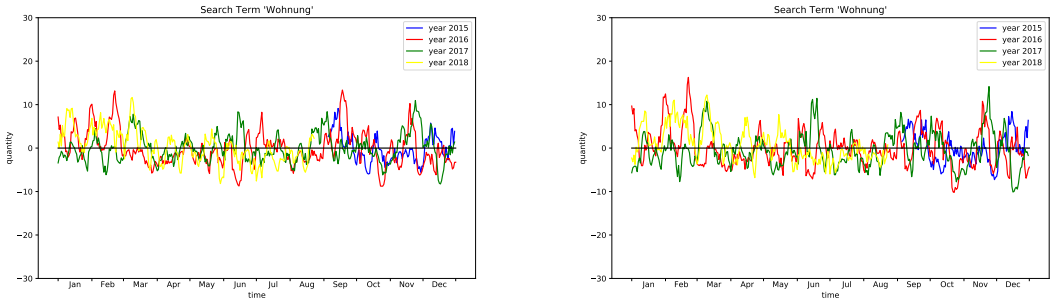

Figure 3: Seasonal overview of the de-trended indices ZH (left) and BE (right) for the years 2015-2018. The x-axis describes the days in a year, and the y-axis the daily popularity of the search term "Wohnung". Source: own elaboration based on Google Trends (2018).

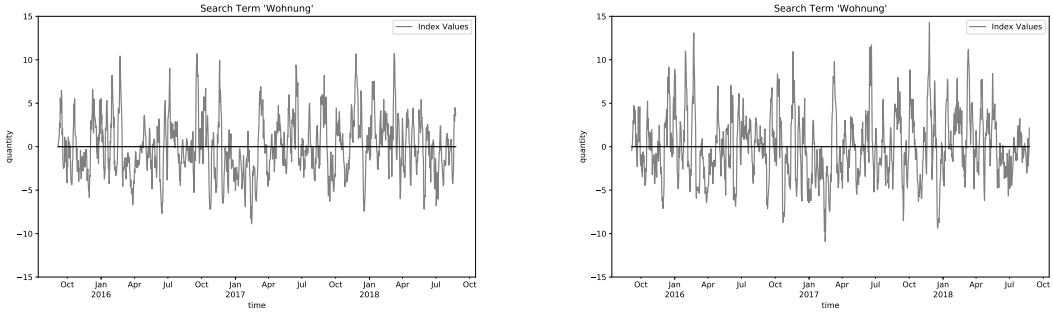

Figure 4: Plots of the de-trended and de-seasonalized indices ZH (left) and BE (right). The x-axis describes the time from 6 September 2015 to 24 August 2018, and the y-axis the daily popularity of the search term "Wohnung". Source: own elaboration based on Google Trends (2018).

p/q	0	1	2	3	4	5	6
0	0.86	0.711	0.572	0.425	0.285	0.138	0.008
1	-0.108	-0.053	-0.094	-0.068	-0.087	-0.065	0.005
2	-0.049	-0.01	0.009	-0.008	0.011	0.034	0.006
3	-0.119	-0.01	0	0	-0.005	-0.017	0.004
4	-0.075	-0.011	0.001	-0	0.003	0.009	0.002
5	-0.141	-0.016	-0.006	-0.003	-0.002	-0.004	0.002
6	-0.057	0.059	-0.018	0.012	-0.004	0.003	0.001

p/q	0	1	2	3	4	5	6
0	0.854	0.709	0.575	0.43	0.29	0.133	-0.018
1	-0.077	-0.04	-0.097	-0.067	-0.1	-0.084	-0.017
2	-0.04	-0.008	0.012	-0.012	0.015	0.041	-0.013
3	-0.129	-0.013	-0	0	-0.007	-0.022	-0.007
4	-0.08	-0.017	0.001	0	0.004	0.011	-0.004
5	-0.179	-0.023	-0.01	-0.005	-0.002	-0.006	-0.002
6	-0.109	0.079	-0.027	0.016	-0.006	0.003	-0

Table 2: Vector-correlation tables for the error series of the indices ZH (top) and BE (bottom). Source: own elaboration based on Google Trends (2018).

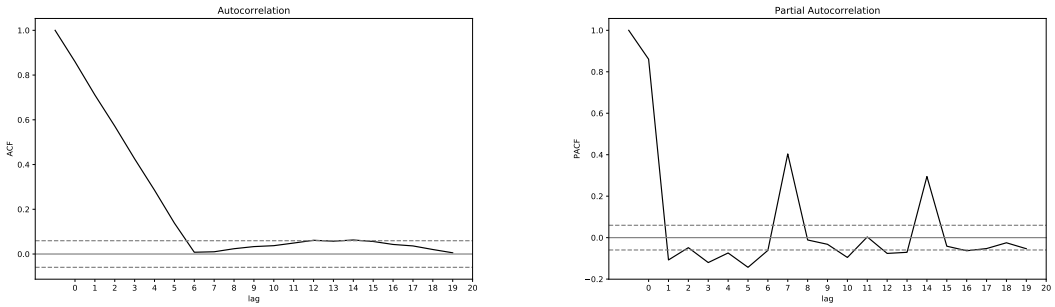

Figure 5: ACF and PACF plots of the de-trended and de-seasonalized index ZH. The x-axis describes the lag, and the y-axis the function value. Source: own elaboration based on Google Trends (2018).

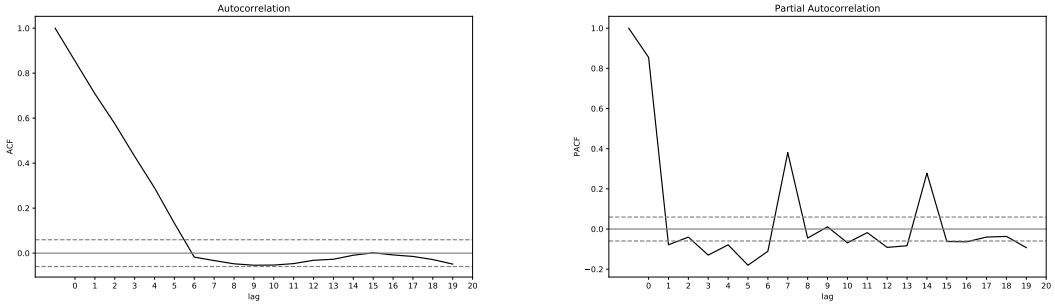

Figure 6: ACF and PACF plots of the de-trended and de-seasonalized index BE. The x-axis describes the lag, and the y-axis the function value. Source: own elaboration based on Google Trends (2018).

p/q	1	2	3	4	5	6	7
1	4235	4235.92	4230.63	4180.78	4125.25	3838.01	3839.38
2	4219.52	4220.41	4158.86	4115.3	4044.77	3838.55	3841.76
3	4220.8	4223.33	4098.02	4086.35	4002.57	3840.43	3842.55
4	4211.8	4089.26	4126.18	4085.84	3976.75	3842.17	3843.96
5	4211.27	4205.58	4086.68	4003.11	3964.65	3844.17	3842.96
6	4196.95	4082.14	4065.5	3966.58	3955.75	3844.79	3848.12
7	4128.98	4035.48	4010.57	3981.53	3967.42	3846.26	3848.3

p/q	1	2	3	4	5	6	7
1	4568.38	4569.78	4562.9	4512.27	4444.16	4118.48	4120.24
2	4549.01	4548.37	4497.86	4449.54	4357.09	4120.5	4120.59
3	4549.32	4550.55	4420.25	4405.36	4307.72	4122.48	4122.59
4	4537.57	4409.78	4444.16	4408.47	4285.78	4124.5	4125.89
5	4536.6	4530.44	4405.43	4308.32	4268.51	4126.07	4127.92
6	4513.23	4398.92	4371.47	4271.47	4262.43	4127.76	4129.01
7	4424.37	4349.16	4325.66	4268.54	4272.53	4128.57	4131

Table 3: AICc tables for the error series of the indices ZH (top) and BE (bottom). Source: own elaboration based on Google Trends (2018).

	coeff	pvals	conf_lower	conf_higher
const	-0.002587	9.921541e-01	-0.518062	0.512888
ar.L1.index	0.020760	6.391573e-01	-0.065996	0.107516
ma.L1.index	0.953091	3.692507e-138	0.889008	1.017175
ma.L2.index	0.866583	9.852624e-124	0.803915	0.929251
ma.L3.index	0.848609	3.796745e-142	0.792631	0.904587
ma.L4.index	0.807087	8.094442e-149	0.755478	0.858695
ma.L5.index	0.802105	6.346257e-133	0.746768	0.857442
ma.L6.index	0.773578	4.550501e-154	0.725274	0.821881

	coeff	pvals	conf_lower	conf_higher
const	-0.001639	9.956706e-01	-0.593571	0.590293
ar.L1.index	-0.023437	5.330013e-01	-0.097095	0.050221
ma.L1.index	0.983989	3.759462e-242	0.940069	1.027908
ma.L2.index	0.909725	2.830169e-187	0.860268	0.959182
ma.L3.index	0.909582	2.807472e-192	0.861082	0.958082
ma.L4.index	0.874935	1.005292e-186	0.827267	0.922604
ma.L5.index	0.868302	5.568239e-206	0.824326	0.912278
ma.L6.index	0.839923	3.503657e-224	0.800066	0.879779

Table 4: Overview of the fitted error models for the indices ZH (top) and BE (bottom). Source: own elaboration based on Google Trends (2018).

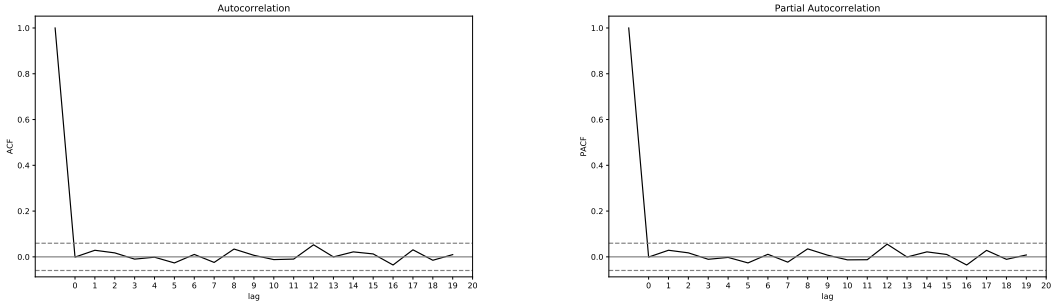

Figure 7: ACF and PACF plots of the error residuals of the index ZH. The x-axis describes the lag, and the y-axis the function value. Source: own elaboration based on Google Trends (2018).

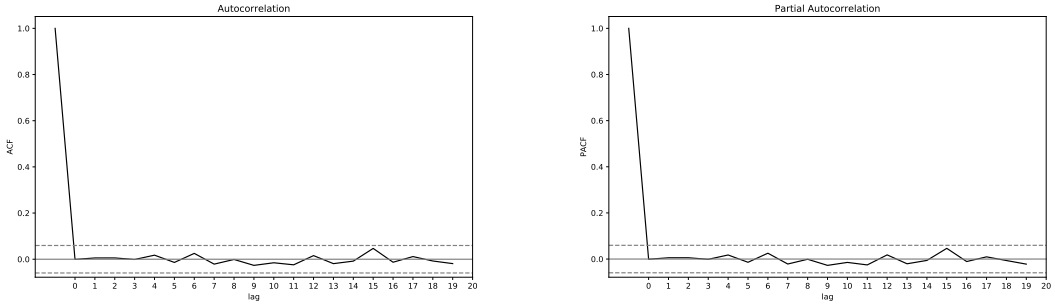

Figure 8: ACF and PACF plots of the error residuals of the index BE. The x-axis describes the lag, and the y-axis the function value. Source: own elaboration based on Google Trends (2018).

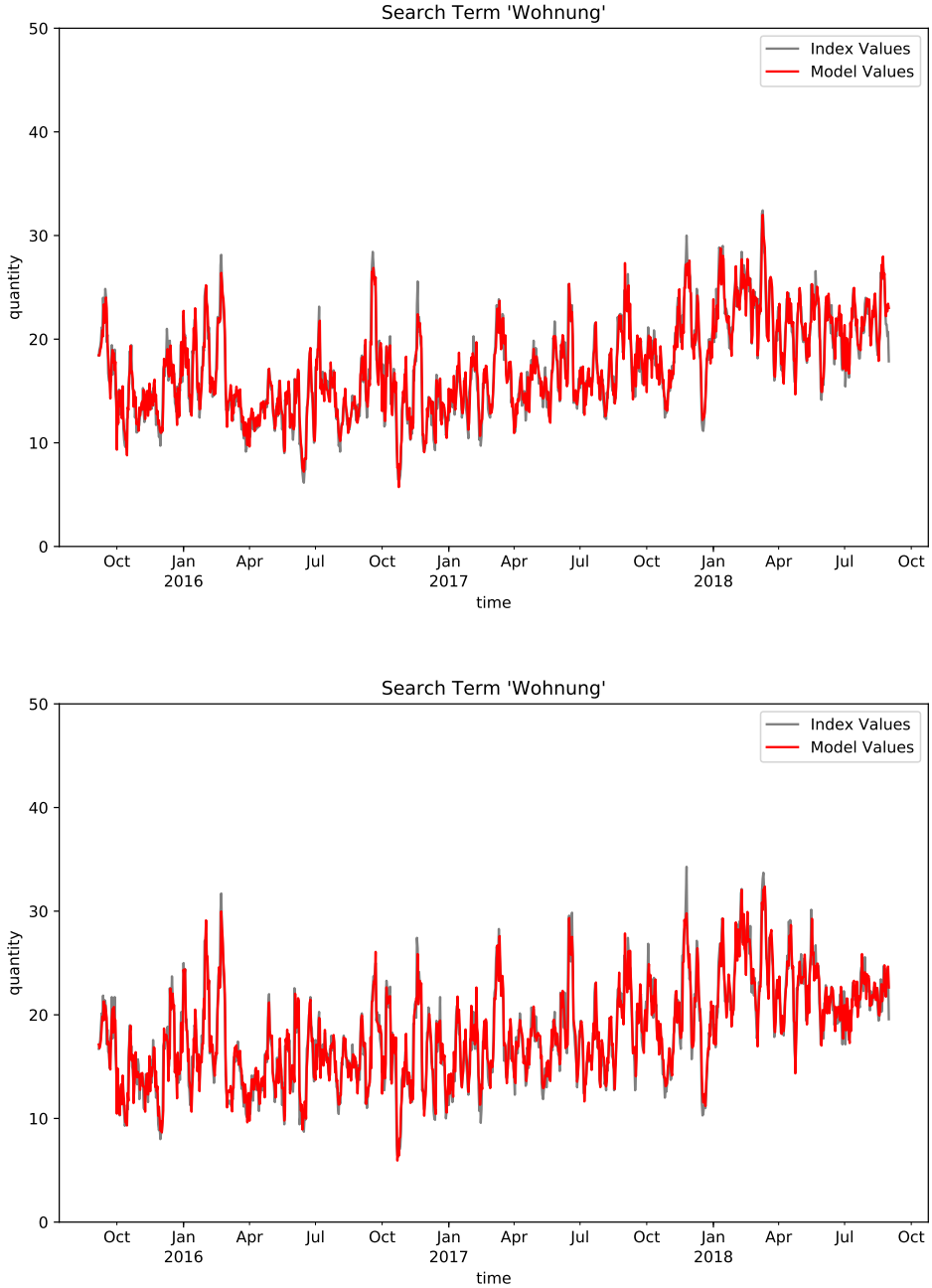

Figure 9: Plots of the indices ZH (top) and BE (bottom) with their decomposition model functions. The x-axis describes the time from 6 September 2015 to 31 August 2018, and the y-axis the daily popularity of the search term "Wohnung". Source: own elaboration based on Google Trends (2018).

References

Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer, New York, 23 edition, 1991.

Google Trends. Google Trends Erkunden, 2018. URL `https://trends.google.ch/trends/explore?q=Wohnung&geo=CH`. Accessed: 2018-10-07.

Rainer Schlittgen. *Angewandte Zeitreihenanalyse*. Oldenbourg Wissenschaftsverlag, München, 2001.

Rainer Schlittgen and Bernd H.J. Streitberg. *Zeitreihenanalyse*. Oldenbourg Wissenschaftsverlag, München, 2001.

Lynn Wu and Erik Brynjolfsson. *The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales*, chapter 1, pages 89–118. University of Chicago Press, 2015.