Issue 2

hungary_chickenpox.csv is a spatio-temporal dataset of weekly chickenpox cases from Hungary. The dataset consists of a county-level adjacency matrix and a time series of county-level reported cases between 2005 and 2015. Use an appropriate method to create a forecasting model for Budapest chickenpox cases using data until Jan 6, 2014. Test the accuracy using the rest of the data points.

Answer:

VAR Method

The VAR method, displayed lower AIC value of 122.496 and an impressively low p-value of 0.006 for the forcasting method. This is the best fitted model for dataset.

1. **Introduction**

   Varicella-zoster, highly contagious illness known as chickenpox, exceeds age boundaries, particularly affecting children. In Hungary, the Hungarian National Epidemiology Center carefully accumulates weekly reports of chickenpox cases, affording vital visions into the disease's distribution.

   Within the limits of this research effort, we are prepared with an extensive dataset surrounding a county-level adjacency matrix and a time series, documenting county-level reports of cases crossing the years 2005 to 2015. Our principal research objective revolves around the development of a precise forecasting model, exactly designer to predict chickenpox cases totally within Budapest. To fulfill this aim, we connect data available until January 6, 2014, followed by a valuation of the model's predictive accuracy when deployed on the following dataset.

This data by the National Epidemiology Center of Hungary, has been made accessible to the wider scientific community through the esteemed UCI Machine Learning Repository. Thus, it serves as a invaluable resource for this research, offering a promising avenue to advance our understanding of chickenpox dynamics in Budapest and bolster public health strategies.

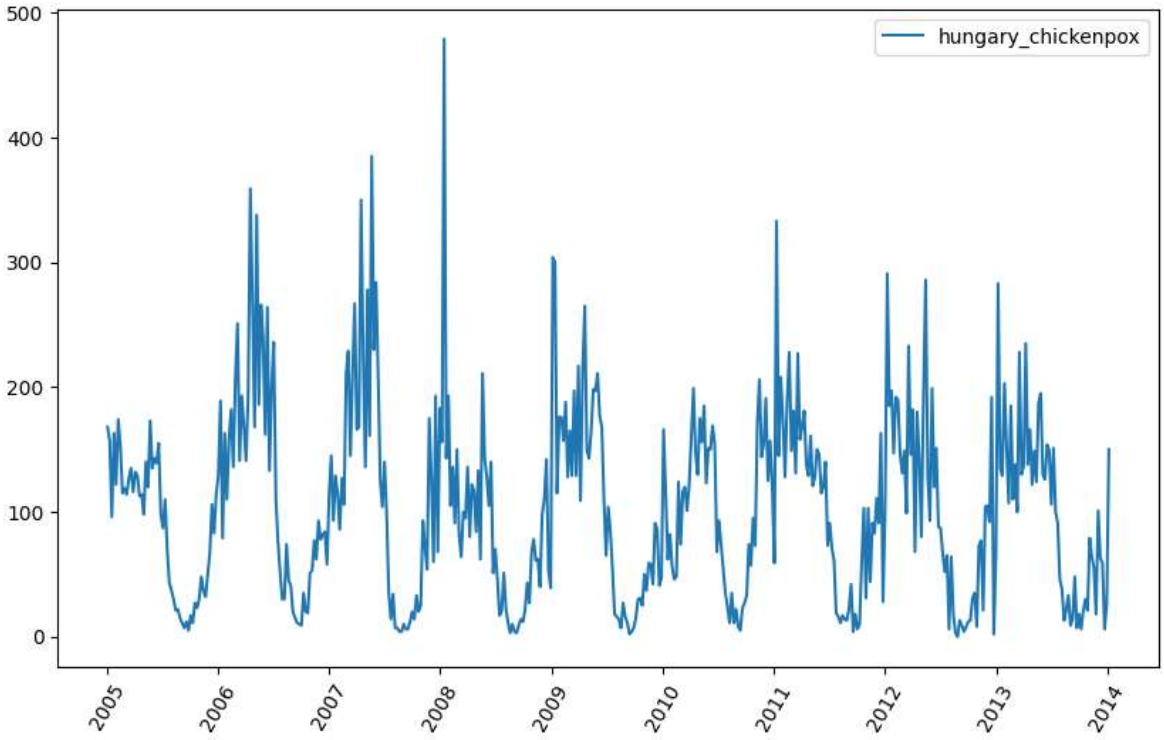
2. Tools

The analysis was conducted using Python in combination with several prominent machine learning frameworks and libraries, including numpy, pandas, matplotlib, seaborn, statsmodels and scikit-learn. For an interactive and collaborative development environment, Anaconda's Jupyter Notebook served as the primary platform for this analysis.

3. Preliminary Data Analysis and Visualization

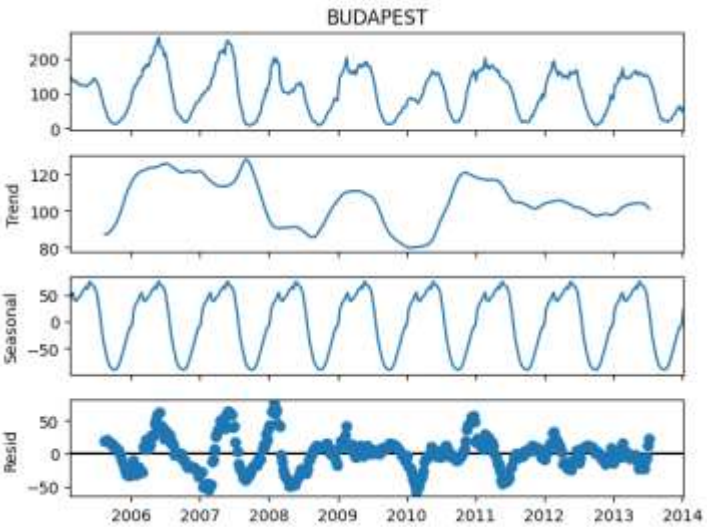
The dataset comprises 520 rows and 21 columns, which include various attributes like Date, Budapest, Baranya, Bacs, Bekes, Borsod, Csongrad, Fejer, Gyor, Hajdu, Heves, Jasz, Komarom, Nograd, Pest, Somogy, Szabolcs, Tolna, Vas, Veszprem, and Zala. Out of these, 470 rows are designated as the training data, with the remaining 50 rows serving as test data. The primary focus of our analysis centers around Budapest, which is regarded as the pivotal feature in the quest to construct a time series model.

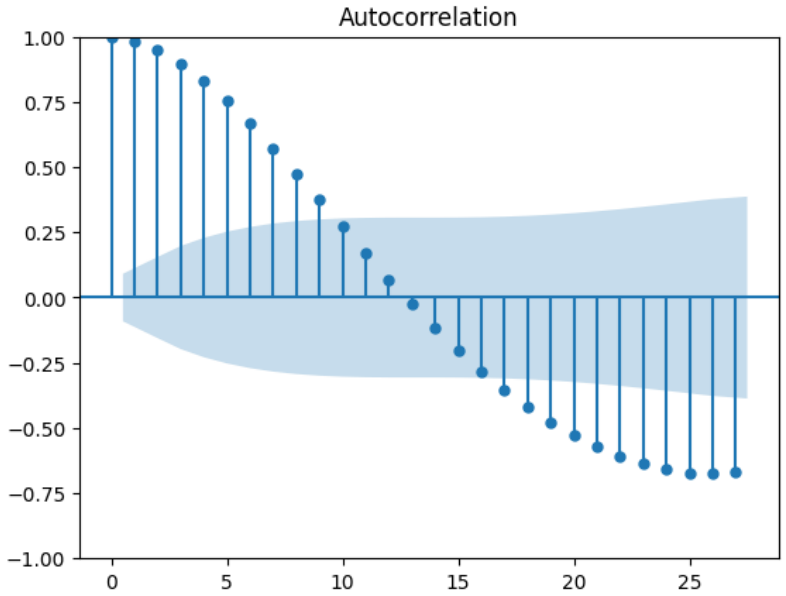
Our central aim is to find an optimal model that can facilitate accurate predictions. This task requires a complete understanding of the dataset. To board on this journey, our initial step involves the creation of a time series graph for Budapest, spanning the time frame from January 3, 2005, to January 6, 2014. This foundational visualization will provide crucial visions to inform subsequent modeling efforts.

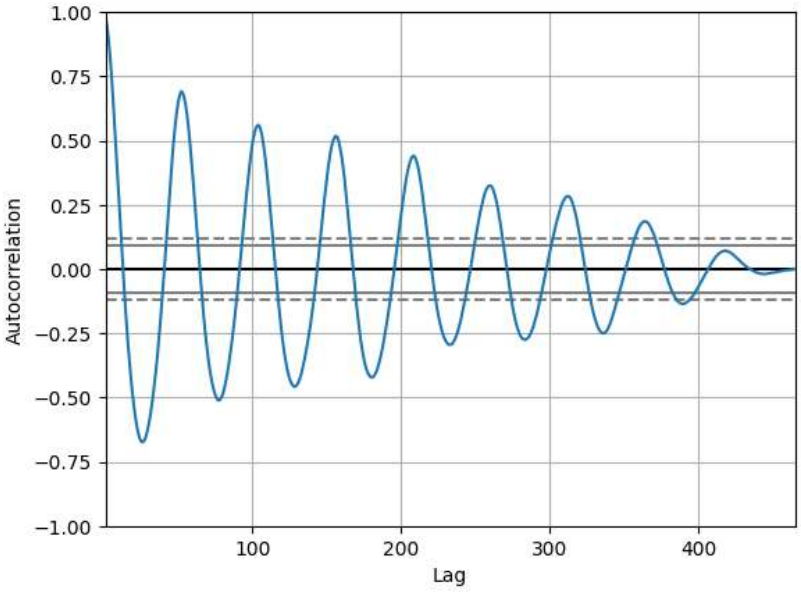


3.1. Correlation analysis and auto correlation

To unveil the underlying patterns in our time series data, we employ the decomposition method. This analysis reveals notable characteristics within the dataset, including significant residual values and pronounced seasonality. As a logical consequence, our approach involves the implementation of data smoothing techniques aimed at reducing the magnitude of these residuals

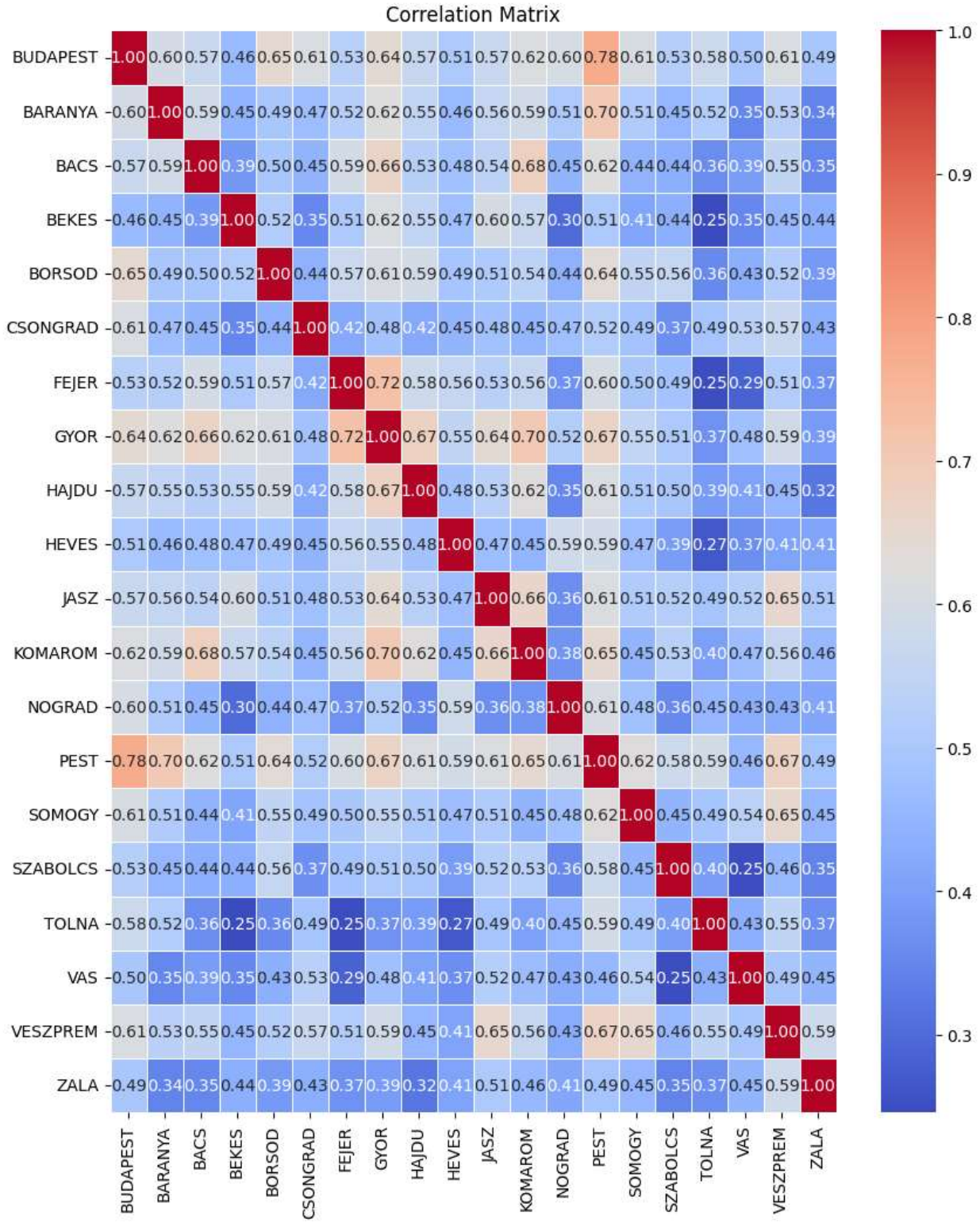






Based on the autocorrelation map, it becomes evident that the trend extends beyond the scope of white noise. Consequently, our next step involves identifying and exploring the correlated

parameters in alignment with the insights derived from the heatmap.



4. Machine Learning techniques used and results

This study leveraged an array of potent analytical techniques, encompassing regression analysis, the difference equation method, equivalent models, vector auto-regression (VAR) models, and second-order difference equation models. The main goal behind the application of these methodologies is their collective operation in addressing the research objectives.

4.1. Regression analysis

The method employed serves as a fundamental model for the analysis of time series data, and it demonstrates a strong fit with our training dataset. With an R-squared value of 0.723 and an adjusted R-squared value of 0.711, it showcases notable goodness-of-fit. The AIC (Akaike Information Criterion) value, standing at 4852, further supports the model's effectiveness. Nevertheless, our pursuit of excellence led us to enhance and refine this model even further.

OLS Regression Results

Dep. Variable:	BUDAPEST	R-squared:	0.723
Model:	OLS	Adj. R-squared:	0.711
Method:	Least Squares	F-statistic:	61.81
Date:	Tue, 24 Oct 2023	Prob (F-statistic):	1.78e-112
Time:	00:42:52	Log-Likelihood:	-2406.2
No. Observations:	471	AIC:	4852.
Df Residuals:	451	BIC:	4936.
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.2517	3.377	3.332	0.001	4.614	17.889
BARANYA	-0.0910	0.089	-1.027	0.305	-0.265	0.083
BACS	0.0524	0.085	0.614	0.539	-0.115	0.220
BEKES	-0.0510	0.072	-0.712	0.477	-0.192	0.090
BORSOD	0.2764	0.056	4.937	0.000	0.166	0.386
CSONGRAD	0.4399	0.078	5.675	0.000	0.288	0.592
FEJER	-0.0829	0.098	-0.843	0.400	-0.276	0.110
GYOR	0.1871	0.104	1.797	0.073	-0.017	0.392
HAJDU	0.0154	0.066	0.234	0.815	-0.114	0.145
HEVES	-0.1589	0.087	-1.820	0.069	-0.330	0.013
JASZ	0.0195	0.086	0.227	0.821	-0.149	0.188
KOMAROM	0.2999	0.134	2.243	0.025	0.037	0.563

4.2. Difference method.

To enhance our model, we incorporated the difference equation model. However, this adjustment yielded a lower R-squared value of 0.308. Additionally, the AIC stood at 5089, and the BIC at 5176. Consequently, it becomes apparent that this model did not perform as effectively as the previous regression model.

OLS Regression Results

Dep. Variable:	Diff	R-squared:	0.308
Model:	OLS	Adj. R-squared:	0.277
Method:	Least Squares	F-statistic:	9.973
Date:	Tue, 24 Oct 2023	Prob (F-statistic):	1.91e-25
Time:	00:47:56	Log-Likelihood:	-2523.5
No. Observations:	470	AIC:	5089.
Df Residuals:	449	BIC:	5176.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-24.1380	4.445	-5.430	0.000	-32.873	-15.403
BUDAPEST	0.7579	0.061	12.389	0.000	0.638	0.878
BARANYA	-0.0560	0.115	-0.487	0.627	-0.282	0.170
BACS	-0.1873	0.111	-1.682	0.093	-0.406	0.032
BEKES	-0.0128	0.094	-0.137	0.891	-0.197	0.171
BORSOD	-0.1232	0.075	-1.648	0.100	-0.270	0.024
CSONGRAD	-0.2567	0.104	-2.462	0.014	-0.461	-0.052
FEJER	0.0102	0.129	0.080	0.937	-0.242	0.263
GYOR	-0.1784	0.136	-1.315	0.189	-0.445	0.088
HAJDU	-0.0679	0.085	-0.794	0.427	-0.236	0.100
HEVES	0.3187	0.114	2.787	0.006	0.094	0.543
JASZ	-0.0455	0.112	-0.407	0.684	-0.265	0.174

4.3 Equivalent model

In comparison to the difference equation model, this equivalent model demonstrates superior performance. However, it falls slightly short of the excellence achieved by the regression model. The equivalent model exhibits an R-squared value of 0.533, alongside AIC and BIC values of 5089 and 5176, respectively. It's worth noting that the AIC and BIC values in this case are relatively higher, indicating that there is room for further improvement.

OLS Regression Results

Dep. Variable:	Shift	R-squared:	0.533
Model:	OLS	Adj. R-squared:	0.513
Method:	Least Squares	F-statistic:	25.67
Date:	Tue, 24 Oct 2023	Prob (F-statistic):	8.04e-62
Time:	00:57:15	Log-Likelihood:	-2523.5
No. Observations:	470	AIC:	5089.
Df Residuals:	449	BIC:	5176.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	24.1380	4.445	5.430	0.000	15.403	32.873
BUDAPEST	0.2421	0.061	3.958	0.000	0.122	0.362
BARANYA	0.0560	0.115	0.487	0.627	-0.170	0.282
BACS	0.1873	0.111	1.682	0.093	-0.032	0.406
BEKES	0.0128	0.094	0.137	0.891	-0.171	0.197
BORSOD	0.1232	0.075	1.648	0.100	-0.024	0.270
CSONGRAD	0.2567	0.104	2.462	0.014	0.052	0.461
FEJER	-0.0102	0.129	-0.080	0.937	-0.263	0.242
GYOR	0.1784	0.136	1.315	0.189	-0.088	0.445

4.4 Vector auto regression model (VAR)

To apply the vector auto-regression model effectively, we divided the dataset into four lags. This strategic division yields significantly improved AIC and BIC values, providing a more accurate representation of the dataset. The comprehensive consideration of all relevant parameters results in an exceptionally well-fitting model, as evidenced by the AIC value of 122.496 and a low p-value of 0.006. This model is a robust fit for our forecasting methodology.

```
Lag Order = 1
AIC :  122.49619176540209
BIC :  126.20714438620149
FPE :  1.584664313343181e+53
HQIC:  123.95617668441008

Lag Order = 2
AIC :  120.3364244848589
BIC :  127.59334190517009
FPE :  1.842177961103099e+52
HQIC:  123.19173822592074

Lag Order = 3
AIC :  120.29864972609697
BIC :  131.11303289066862
FPE :  1.8113177246149097e+52
HQIC:  124.55406704119683

Lag Order = 4
AIC :  120.30725769040646
BIC :  134.69066967632426
```

```
FPE :  1.9034912603092479e+52
HQIC:  125.96757956029293
```

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Tue, 24, Oct, 2023
Time:                      01:00:27
--------------------------------------------------------------------
No. of Equations:         20.0000    BIC:                    126.207
Nobs:                     470.000    HQIC:                   123.956
Log likelihood:          -41704.6    FPE:                 1.58466e+53
AIC:                      122.496    Det(Omega_mle):      6.61096e+52
--------------------------------------------------------------------
Results for equation BUDAPEST
================================================================================
                 coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------------
const              20.997656         4.254382            4.936           0.000
L1.BUDAPEST         0.161958         0.058597            2.764           0.006
L1.BARANYA         -0.088689         0.110375           -0.804           0.422
L1.BACS             0.094053         0.106694            0.882           0.378
L1.BEKES           -0.090635         0.089794           -1.009           0.313
L1.BORSOD          -0.018496         0.072676           -0.254           0.799
L1.CSONGRAD         0.209452         0.099899            2.097           0.036
L1.FEJER           -0.034423         0.122684           -0.281           0.779
L1.GYOR             0.466870         0.130399            3.580           0.000
```

4.5 Testing data with VAR model

After thorough analysis, the Vector Auto-Regression (VAR) method emerged as the most suitable and effective model. Subsequently, we rigorously tested the data using this chosen method.

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Tue, 24, Oct, 2023
Time:                      04:34:16
--------------------------------------------------------------------
No. of Equations:         20.0000    BIC:                    114.565
Nobs:                     49.0000    HQIC:                   104.502
Log likelihood:          -3380.12    FPE:                 1.68430e+43
AIC:                      98.3495    Det(Omega_mle):      1.34394e+40
--------------------------------------------------------------------
Results for equation BUDAPEST
================================================================================
                 coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------------
const             15.901051        15.913012            0.999           0.318
L1.BUDAPEST        0.232628         0.302906            0.768           0.442
L1.BARANYA         2.743744         1.056672            2.597           0.009
L1.BACS            0.010847         0.336407            0.032           0.974
L1.BEKES          -1.125519         1.068251           -1.054           0.292
L1.BORSOD         -0.206801         0.308381           -0.671           0.502
L1.CSONGRAD       -0.736644         0.308305           -2.389           0.017
L1.FEJER           0.502420         0.692393            0.726           0.468
L1.GYOR            1.563345         0.693484            2.254           0.024
L1.HAJDU          -0.028567         0.506741           -0.056           0.955
L1.HEVES           0.045818         0.607577            0.075           0.940
```

5. Results and Conclusions

Certainly, here's the provided information rewritten:

Regression Model

The regression model yielded an R-squared value of 0.723 and an adjusted R-squared value of 0.711, indicating a substantial goodness-of-fit. The AIC (Akaike Information Criterion) value for this model was 4852.

Difference Method

In contrast, the difference method produced a lower R-squared value of 0.308. Additionally, the AIC was 5089, and the BIC stood at 5176.

Equivalent Method

The equivalent method showcased an R-squared value of 0.533. The AIC and BIC values for this model were 5089 and 5176, respectively.

VAR Method

The VAR method, when assessed, displayed a notably lower AIC value of 122.496 and an impressively low p-value of 0.006.