Project Number 288094

eCOMPASS

eCO-friendly urban Multi-modal route PlANning Services for mobile uSers

STREP
Funded by EC, INFSO-G4(ICT for Transport) under FP7

eCOMPASS – TR – 045

Checking the Stationarity of the Datasets

T. Diamantopoulos, D. Kehagias

October 2013

Project Number 288094

eCOMPASS

eCO-friendly urban Multi-modal route PlANning Services for mobile uSers

STREP
Funded by EC, INFSO-G4(ICT for Transport) under FP7

eCOMPASS – TR – 045

Checking the Stationarity of the Datasets

T. Diamantopoulos, D. Kehagias

October 2013

Checking the Stationarity of the Datasets

1 Introduction

This document provides stationarity checks for the Berlin and Warsaw datasets. At first, a time series is defined as stationary if its joint probability distribution does not change if it is shifted in time. Consequently, variables such as its mean and its variance are also expected not to change over time. Formally a time series x is stationary if the distribution function of its probability distribution X is constant over a time lag τ, i.e.:

$$X(x_1, x_2, \ldots, x_t, \ldots, x_T) = X(x_{1+\tau}, x_{2+\tau}, \ldots, x_{t+\tau}, \ldots, x_{T+\tau}) \tag{1}$$

Although equation (1) is an equality, in practice an approximate equality may be enough. In other words, a time series may be *stationary enough* for our requirements given certain thresholds.

With respect to traffic prediction scenarios, one should confirm that the time series of the roads that are used are stationary in order to apply well known time series algorithms (i.e. Space Time Auto-Regressive Integrated Moving Average (STARIMA) or even simple Auto-Regressive Moving Average (ARMA) models). Intuitively, applying a time series model to predict a future value requires the series to be stationary so that it is actually predictable. Note, however, that if the series is trend-stationary (e.g. it has a mean that augments and a rather stable variance), there are suitable transformations to make it stationary.

2 Taking a look into traffic time series

Before checking if the two datasets at hand contain stationary time series, we shall take a peek on some example series to identify the nature of the data.

Concerning the Berlin dataset, Figure 1 depicts four example road time series for an 8-hour set for the city of Berlin. As shown in that figure, the time series indeed seem stationary. Their mean is rather stable, and their deviation is mostly within steady limits. Further analyzing the figure, one could observe that the speed of the roads is generally within reasonable levels. However, there are some "spikes" in the figure for road 16415. In the case of traffic, they may correspond to "noise" in data. Noise is very common in traffic scenarios since e.g. vehicles may stop at any point in order for the drivers to check something, speak on the phone, fill up gas, etc. As far as the speed time series is concerned,

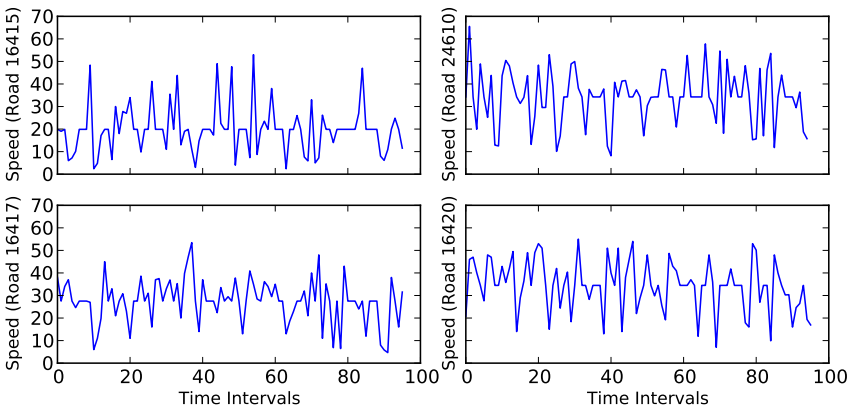

Figure 1: Example time series for four roads of the city of Berlin.

there are several procedures for normalizing their behavior. In our case, having 5-minute intervals and mapping links to roads ensures that these spikes are "absorbed" by the deviation. Thus, upon cleaning the dataset, one could state that real-life traffic scenarios are quite well suited to parametric time series methods. Examples of well performing time series algorithms when applied to real data are given in [1] and [2]. The contrast between real and simulated data (which is also explored in [2]) shall be made clear hereafter.

Concerning the Warsaw dataset, Figure 2 depicts four example road time series for a 10-hour set for the city of Warsaw. As shown in that figure, the series generally suffer from noise in the data. Since the dataset is simulated [3], this noise is actually deliberate, possibly in order to make the problem harder. Thus, feature selection methods such as the one given in [4] are suited to the dataset since they handle noisy data satisfactorily (e.g. taking the number of zero samples into account). Concerning time series methods, such as STARIMA, they are obviously expected to perform somewhat worse [2]. As in the Berlin dataset, a leap from links to roads, as well as the use of time intervals are utilized in order to ensure that the data is "cleaner", with as few spikes as possible. Concerning stationarity, certain series may seem stationary, such as the one of road 9571, whereas others may are clearly non-stationary, such as the one of road 142 which is full of spikes.

A first look into the data indeed provided useful intuition as to the different nature of the datasets. Concerning stationarity, the time series of the Berlin dataset seem to be stationary, whereas the ones of the Warsaw dataset are in question. In any case, the stationarity of the series is properly tested in the following section to support these remarks.

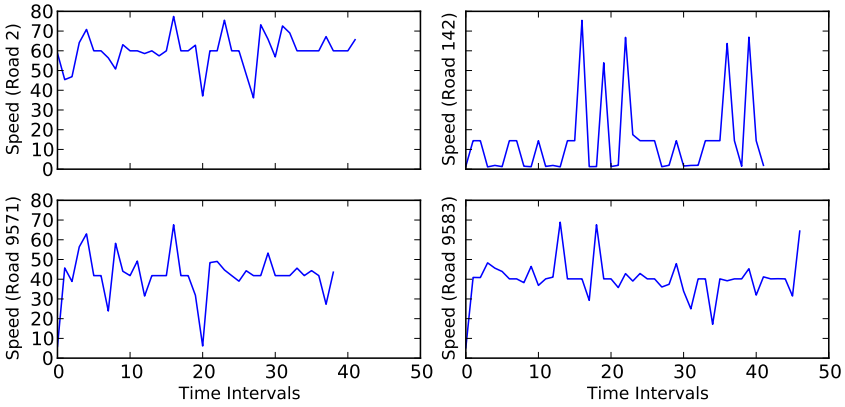

Figure 2: Example time series for four roads of the city of Warsaw.

3 Checking the stationarity

Concerning traffic scenarios, one should test separately whether each of the road time series are stationary, as in [1]. Thus, the task is reduced to checking the stationarity of a time series. This is a well known problem in the literature [5, 6, 7, 8, 9]. In our case, we selected the *Augmented Dickey Fuller (ADF)* test [6] to test the stationarity of the time series. The purpose of the test is to detect the (in)existence of *unit roots* in the time series. A unit root is an element of the series that affects its trend. Intuitively, a time series having a unit root appears to augment over time within small time periods. Note, however, that having a fixed trend is not necessarily due to a unit root. The time series may just be trend-stationary. The root could result in an unexpected trend that may or may not wear off over time but will always start from a specific point in time.

Concerning the ADF test, it is a generalization of the *Dickey Fuller (DF)* test [5] which is applied to the model:

$$\Delta x_t = \beta_0 + \beta_1 t + (\alpha - 1)x_{t-1} + \epsilon_t \tag{2}$$

where x_t is the tested series. β_0 and β_1 are used to allow the series to have a linear trend[1]. The α parameter tests for the existence of a unit root for lag 1. Checking for different lag values involves using the ADF given as follows:

$$\Delta x_t = \beta_0 + \beta_1 t + (\alpha - 1)x_{t-1} + \gamma_1 \Delta x_{t-1} + \gamma_2 \Delta x_{t-2} + \cdots + \gamma_{p-1} \Delta x_{t-p+1} + \epsilon_t \tag{3}$$

Equations (2) and (3) are similar. The γ terms of equation (3) introduce the notion of lag. Thus, the unit roots are sought over the differentiated series from

[1] As noted also in [8], one can use similar equations to (2) in order to impose different restrictions. For example, if β_1 is 0 then the series is only allowed to have a non-zero mean but not a trend. Similarly, one could also add a β_2 to allow for a non-linear trend, which is actually rarely used.

lag 1 to lag p. The null hypothesis for both equations is $\alpha = 1$, i.e. that there is a unit root. Intuitively, equation (3) should be almost constant with respect to new time series values.

Upon further research [8, 9], the p-values for the test are obtained using regression surface approximation. Thus, the test provides an approximate p-value that represents how much possible is that the series has a unit root. For example, if the p-value is lower than 0.005 then we can be sure with confidence 0.5% that the time series has no unit roots.

4 Results

We performed the ADF test for all the road time series for each road network. We tested an 8-hour set for Berlin and a 10-hour set for Warsaw. Both sets were tested for the inexistence of unit roots when no constant was used ($\beta_0 = 0$ and $\beta_1 = 0$), a constant for the mean of the series was used ($\beta_0 \neq 0$ and $\beta_1 = 0$), and two constants for the mean and the trend of the mean of the series were used ($\beta_0 \neq 0$ and $\beta_1 \neq 0$). The results for Berlin are shown in Table 1.

Table 1: Percentage of Stationary Road Time Series for Berlin

Confidence	**Roads (%)**		
	$\boldsymbol{\beta_0 = 0, \beta_1 = 0}$	$\boldsymbol{\beta_0 \neq 0, \beta_1 = 0}$	$\boldsymbol{\beta_0 \neq 0, \beta_1 \neq 0}$
0.1%	0.03%	88.96%	90.51%
0.5%	0.03%	2.59%	1.70%
1%	0.03%	1.17%	0.96%
5%	1.83%	1.68%	1.80%
10%	4.44%	0.66%	0.63%
100%	93.65%	2.41%	2.56%

in Table 1, almost 90% of the road time series of the dataset are stationary around a constant mean (β_0) with confidence below 0.1%. Thus, we can fairly claim that time series algorithms can be used on this dataset. As expected, when the trend of the mean is also taken into account (β_1), the percentage of stationary series is even higher. On the other hand, not using any constant provides unsatisfactory results. Intuitively, this is expected since the time series of a road usually has a constant mean speed value that mainly depends on the size of the road. E.g. a ring road could have mean speed around 100 km/h with little variation. Obviously, having a zero mean requires having many near-to-zero values since speed has no negative values, which is rather rare. Having a time series with speed values that constantly increase is also rather rare.

Concerning the Warsaw dataset, the results are quite different. Since the dataset is simulated, the stationarity criterion is actually dependent on the decisions made when creating the scenario. The results are shown in Table 2.

Table 2: Percentage of Stationary Road Time Series for Warsaw

Confidence	**Roads (%)**		
	$\boldsymbol{\beta_0 = 0, \beta_1 = 0}$	$\boldsymbol{\beta_0 \neq 0, \beta_1 = 0}$	$\boldsymbol{\beta_0 \neq 0, \beta_1 \neq 0}$
0.1%	15.67%	46.22%	43.79%
0.5%	4.68%	7.34%	7.78%
1%	2.29%	3.35%	3.72%
5%	8.22%	8.63%	8.66%
10%	5.27%	4.02%	3.91%
100%	60.27%	21.27%	22.23%

As shown in Table 2, there are more non-stationary road time series in the Warsaw dataset than in the Berlin dataset. However, when no constant is used, more than 20% of the time series are stationary with confidence 0.5%. These series actually have zero means, and regarding traffic this means that they have speed values marginally over zero. As one can observe, when a constant mean is taken into account, more than more than 53% of the time series are stationary with confidence 0.5%. In addition, more than 65% of the series are stationary with confidence 5%. These stats roughly indicate that applying a time series algorithm is possible, yet not optimal. Concerning the use of a trend parameter, the stationary series are somewhat fewer, probably because the data do not have any actual trend.

5 Conclusion

The stationarity of the time series is an important criterion that cannot be taken for granted in any scenario. Concerning time series corresponding to real traffic data, our intuitive analysis showed that the series are indeed expected to be stationary. In any case, we did confirm (as we had to) that the level of stationarity is high. Concerning simulated datasets, the stationarity criterion is obviously dependent on the decisions made by the creator(s) of the dataset.

Having a stationary time series means that a time series model is expected to provide satisfactory results. When the percentage of the dataset is not very high, one can choose either to apply the algorithms, yet understanding that they may not be optimal, or to transform the series to achieve stationarity. Thus, future work may also lie in applying these transformations.

References

[1] Yiannis Kamarianakis and Poulicos Prastacos. Space-time modeling of traffic flow. *Comput. Geosci.*, 31(2):119–133, March 2005.

[2] Themistoklis Diamantopoulos, Dionysios Kehagias, Felix G. König, and Dimitrios Tzovaras. Investigating the effect of global metrics in travel time forecasting. To appear in *Proceedings of 16th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2013.

[3] Marcin Wojnarski, Pawel Gora, Marcin Szczuka, Hung Son Nguyen, Joanna Swietlicka, and Demetris Zeinalipour. IEEE ICDM 2010 contest: Tomtom traffic prediction for intelligent GPS navigation. *2012 IEEE 12th International Conference on Data Mining Workshops*, 0:1372–1376, 2010.

[4] Benjamin Hamner. Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, pages 1357–1359, Washington, DC, USA, 2010. IEEE Computer Society.

[5] David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431, 1979.

[6] Said E. Said and David A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607, 1984.

[7] Peter C. B. Phillips and Pierre Perron. Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346, 1988.

[8] James G. MacKinnon. Approximate asymptotic distribution functions for unit roots and cointegration tests. Working Papers 861, Queen's University, Department of Economics, November 1992.

[9] James G. MacKinnon. Critical values for cointegration tests. Working Papers 1227, Queen's University, Department of Economics, January 2010.

[10] G. E. P. Box and D. R. Cox. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

[11] R. M. Sakia. The Box-Cox Transformation Technique: A Review. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(2):169–178, January 1992.

[12] Jason W Osborne. Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research & Evaluation*, 15(12):2, 2010.

A Appendix

Since the results for the city of Warsaw were not very satisfactory, we attempted to improve them by transforming the time series. We used the Box-Cox power transformation [10] since it highly supported by literature on the subject [11, 12]. The method transforms series x_t to a new series x'_t as follows:

$$x'_t = \begin{cases} \frac{x_t^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ log(x_t) & \text{if } \lambda = 0 \end{cases} \tag{4}$$

where λ is a parameter that defines the transformation. Note also that the transformation is reversible. Concerning the λ parameter, different values correspond to different transformations[2]. We performed the transformations and checked the stationarity at level 0.1% for different values of λ from -3 to 1 with step 0.25. Our results are shown in Table 3.

Table 3: Stationary Road Time Series Percentage for different λ values at 0.1%

Box-Cox λ	**Roads (%)**		
	$\beta_0 = 0, \beta_1 = 0$	$\beta_0 \neq 0, \beta_1 = 0$	$\beta_0 \neq 0, \beta_1 \neq 0$
-3	1.62%	45.37%	42.09%
-2.8	1.81%	45.19%	41.80%
-2.6	2.14%	44.71%	41.67%
-2.4	2.29%	44.43%	40.97%
-2.2	2.58%	44.01%	40.88%
-2	2.80%	44.23%	40.62%
-1.8	3.06%	44.51%	40.82%
-1.6	3.54%	44.47%	41.28%
-1.4	3.84%	44.30%	40.79%
-1.2	4.50%	44.65%	40.27%
-1	5.35%	44.42%	40.60%
-0.8	5.90%	44.85%	41.23%
-0.6	7.08%	44.74%	40.95%
-0.4	8.55%	44.38%	40.32%
-0.2	8.89%	44.66%	40.60%
0	10.15%	44.82%	41.60%
0.2	12.43%	46.11%	41.36%
0.4	14.65%	46.00%	41.54%
0.6	17.35%	45.21%	41.91%
0.8	19.23%	44.80%	41.80%
1	21.13%	46.20%	43.70%

[2]E.g., as mentioned in [12], setting it to 0.5 provides the square rooted series, setting it to 0.25 provides the fourth rooted series, setting it to 0 provides the logarithm, setting it to -0.5 provides the reciprocal square rooted series, etc.

As one can observe, the transformation did not achieve to improve on the data. The results for confidence level 0.5% are shown in Table 4.

Table 4: Stationary Road Time Series Percentage for different λ values at 0.5%

Box-Cox λ	**Roads (%)**		
	$\beta_0 = 0, \beta_1 = 0$	$\beta_0 \neq 0, \beta_1 = 0$	$\beta_0 \neq 0, \beta_1 \neq 0$
-3	2.51%	8.74%	9.80%
-2.8	2.73%	8.55%	10.03%
-2.6	2.91%	8.93%	10.24%
-2.4	3.14%	8.92%	10.36%
-2.2	3.24%	8.99%	10.58%
-2	4.02%	8.88%	10.50%
-1.8	4.50%	8.48%	10.36%
-1.6	4.61%	8.48%	10.39%
-1.4	5.09%	8.30%	10.32%
-1.2	4.90%	8.01%	10.46%
-1	4.68%	7.67%	10.46%
-0.8	4.68%	7.86%	10.10%
-0.6	4.75%	8.27%	9.73%
-0.4	5.09%	7.56%	10.14%
-0.2	5.39%	7.30%	9.44%
0	5.39%	7.41%	8.51%
0.2	5.57%	7.34%	8.63%
0.4	5.54%	6.97%	8.85%
0.6	5.45%	6.74%	8.74%
0.8	5.56%	7.30%	8.63%
1	5.05%	7.33%	7.81%

As shown in both of the above tables, the BoxCox transformation was not enough to make the time series stationary.