Physica A 390 (2011) 1677–1683

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Study of cross-correlation in a self-affine time series of taxi accidents

G.F. Zebende [a,b,*], P.A. da Silva [a], A. Machado Filho [a]

[a] Computational Modelling Program -SENAI CIMATEC 41650-010 Salvador, Bahia, Brazil

[b] Physics Department - UEFS 44036-900 Feira de Santana, Bahia, Brazil

ARTICLE INFO

Article history:
Received 2 March 2010
Received in revised form 18 November 2010
Available online 20 January 2011

Keywords:
Time series
Cross-correlation
DFA
DCCA

ABSTRACT

We study in this paper the cross-correlation between self-affine time series of real variables recorded simultaneously in cases of taxi accidents. For this purpose, we apply the DCCA method and show that the cross-correlation can be divided into three distinct groups, if we look for the detrended covariance function, i.e., long-range cross-correlations, short-range cross-correlations and no cross-correlations. Finally, it will be seen that the detrended covariance function is robust, if compared with other methods, in identifying these types of cross-correlations.

1. Introduction

Many time series exhibit complex behavior characterized by long-range power-law correlations [1–4]. These time series can be observed using time records or series of observations. In order to study time series, the autocorrelation function is a possible method for time series analysis. We know that the autocorrelation is a measure that tells how much the value of a realization is able to influence its neighbors. However, a lot of real time series are nonstationary, thus the mean, standard deviation (*sd*), and higher moments, or the autocorrelation functions are not invariant under time translation [5–7]. Nonstationarity, an important aspect of complex variability, can often be associated with different trends in the signal or heterogeneous segments (patches) with different local statistical properties. To address this problem, detrended fluctuation analysis (DFA) was developed to accurately quantify long-range power-law correlations embedded in a nonstationary time series [8,9]. This method provides a single quantitative parameter, the scaling exponent α, to quantify the correlation properties of a signal. One advantage of the DFA method is that it allows the detection of long-range power-law correlations in noisy signals with embedded polynomial trends that can mask the true correlations in the fluctuations of a signal [7,10]. This method has been successfully applied to a wide range of simulated and real time series [11–14]. However, there are cases where many variables are recorded simultaneously, forming time series of equal length N (number of points recorded). These time series can be cross-correlated, as in [15,16] or [17–20], and in this case there are different methods for cross-correlation analysis [21–31]. One of the most recent methods to investigate long-range cross-correlation is detrended cross-correlation analysis (DCCA) [32], a generalization of the DFA method, briefly presented in Section 3.

2. Data

We know that a taxi is a vehicle with a very high daily use, compared to a private vehicle, and this fact is one of reasons why insurance companies charge more for insurance, called the premium, causing them to be much more expensive than

* Corresponding author at: Computational Modelling Program - SENAI CIMATEC 41650-010 Salvador, Bahia, Brazil.
E-mail address: gfzebende@hotmail.com (G.F. Zebende).

0378-4371
doi:10.1016/j.physa.2010.12.038

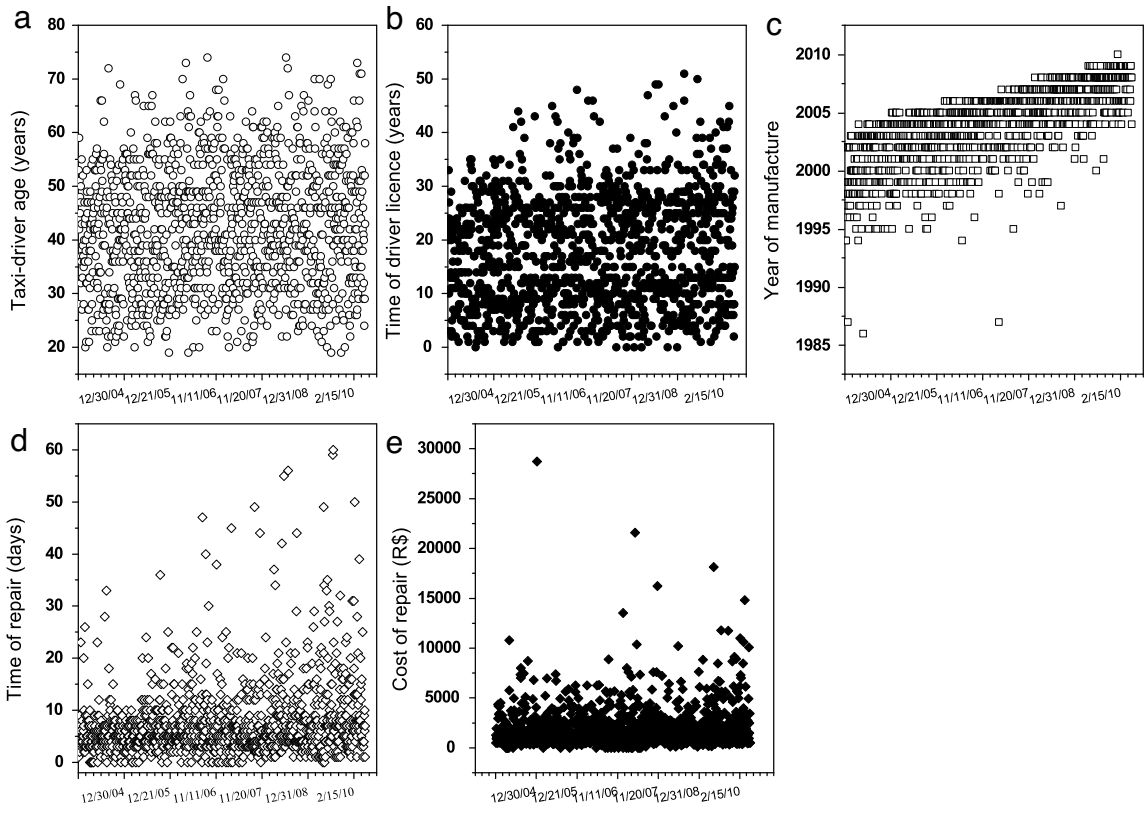

Fig. 1. Original time series of taxi accidents recorded by the CEAT for: (a) (○) taxi driver age, (b) (●) time of driver licence, (c) (□) year of taxi manufacture, (d) (◊) time of taxi repair, and (e) (⧫) cost of taxi repair (in this data range 1 US$ ≅ 1.70 R$). These data correspond to occurrences collected between Aug. 18, 2002 and May 21, 2010, with N = 1250.

private vehicles. Thus, it is better that taxi drivers form cooperatives, because a cooperative, unlike an insurance company, does not take a profit. In this sense, our objective in this paper is to study the cross-correlation between the time series of taxi accidents recorded simultaneously by the CEAT, the special centre for taxi driver support (taxi cooperative with ∼575 associates), located in the city of Salvador, Bahia (Brazil).

Thus, for each accident, a report is made by the CEAT, where several variables are simultaneously recorded. In this paper we analyse five: (a) taxi driver age, (b) time of driver licence, (c) year of taxi manufacture, (d) time of taxi repair, and (e) cost of taxi repair. The data were collected between August 18, 2002 and May 21, 2010 (see Fig. 1). For a more complete visualization we present in Fig. 2 the histogram for these nonstationary time series (not all variables follow a normal distribution). As we all know, a car accident is an unpleasant and unexpected event, which can be generated for different reasons, such as the negligence and carelessness of the driver, problems in the road, and many others. In this way, we can assume that the car accident was generated by a random occurrence. In order to verify if this statement is true or not, in this paper we study the the autocorrelation and the cross-correlation by the DFA and DCCA methods. A brief description of the DFA and the DCCA methods are presented in the section below.

3. The method

The DFA method [8,9] was proposed to analyse long-range correlations in nonstationary time series and provide a relationship between $F_{DFA}(n)$ (root mean square fluctuation function) and the box size n. If there is long-range correlation, then $F_{DFA}(n) \propto n^{\alpha}$. In this way, α is the self-affine scaling exponent such that if $\alpha = 0.50$ the signal is uncorrelated, if $\alpha < 0.50$ the correlation in the signal is antipersistent, and if $\alpha > 0.50$ the correlation in the signal is persistent. Several applications have been made via DFA [6–14].

DCCA [32] is a generalization of the DFA method and is based on detrended covariance; it has many applications [33,34]. This method is designed to investigate power-law cross-correlations between different simultaneously recorded time series in the presence of nonstationarity. They consider two long-range cross-correlated time series $\{y_i\}$ and $\{y'_i\}$ of equal length N, compute two integrated signals $R_k \equiv \sum_{i=1}^{k} y_i$ and $R'_k \equiv \sum_{i=1}^{k} y'_i$, where $k = 1, \ldots, N$. Next we divide the entire time series into $N - n$ overlapping boxes, each containing $n + 1$ values. For both time series, in each box that starts at i and ends at $i + n$, we define the local trend, $\widetilde{R}_{k,i}$ and $\widetilde{R}'_{k,i} (i \leq k \leq i + n)$, to be the ordinate of a linear least-squares fit. We define

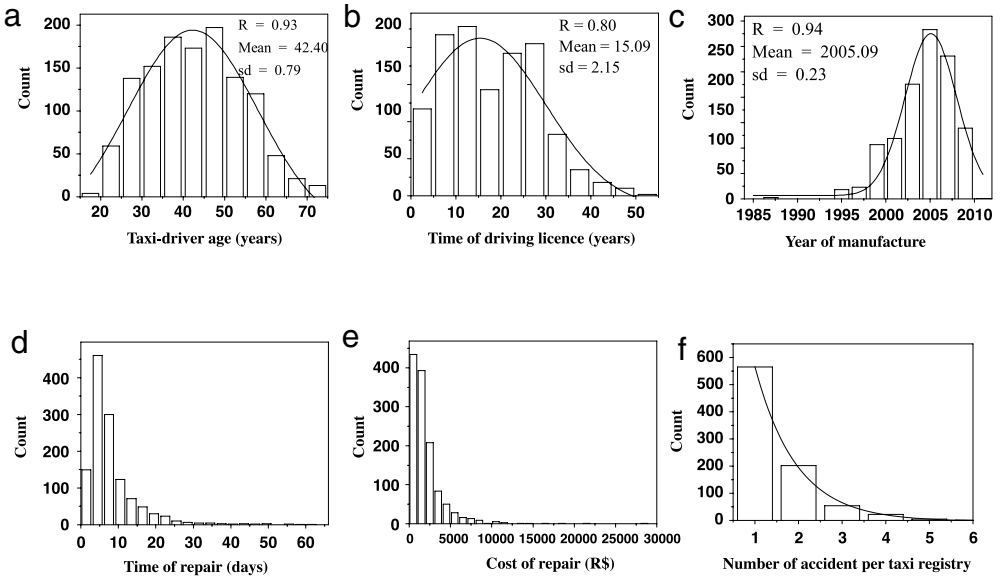

Fig. 2. Histograms for: (a) taxi driver age, (b) time of driver licence, (c) year of taxi manufacture, (d) time of taxi repair, and (e) cost of taxi repair. Continuous lines in (a), (b), and (c) correspond to a Gaussian fit with adjusted R-square, mean and standard error. Cases (d) (time of taxi repair) and (e) (cost of taxi repair) do not fit a Gaussian curve. In this figure, (f) represents the histogram for number of accidents per taxi registry (rank of accidents), and the continuous line represents an exponential decay (only a guide for the eyes).

the detrended walk as the difference between the original walk and the local trend. Next we calculate the covariance of the residuals in each box $f^2_{DCCA}(n,i) \equiv 1/(n+1)\sum_{k=i}^{i+n}(R_k - \widetilde{R}_{k,i})(R'_k - \widetilde{R}'_{k,i})$. Finally, we calculate the detrended covariance function by summing over all overlapping $N-n$ boxes of size n,

$$F^2_{DCCA}(n) \equiv (N-n)^{-1}\sum_{i=1}^{N-n} f^2_{DCCA}(n,i). \tag{1}$$

When only one random walk is analysed ($R_k = R'_k$), the detrended covariance $F^2_{DCCA}(n)$ reduces to the detrended variance $F^2_{DFA}(n)$ used in the DFA method. If self-affinity appears, then $F^2_{DCCA}(n) \sim n^{2\lambda}$.

When $F^2_{DCCA}(n)$ is negative for every n, one may present $-F^2_{DCCA}(n) \times n$ in a log–log plot. Jun et al. [35] proposed a detrended cross-correlation approach to quantify the correlations between positive and negative fluctuations in a single time series, and applied their approach to physiological and financial time series. Zebende and Machado Filho [36] show that it is possible to identify seasonal components with the DCCA method.

4. Results and conclusions

First, in order to identify long-range autocorrelations, we apply the DFA method for the variables recorded simultaneously in taxi accidents by the CEAT (see Fig. 3). In this figure we can see that $F_{DFA}(n) \propto n^{\alpha}$, and the signal is very close to $\alpha = 0.50$ (uncorrelated time series) for taxi driver age, time of driver licence, time of taxi repair, and cost of taxi repair. While for the year of taxi manufacture, the value of α is 0.67 (this data identifies the renewal rate of the taxi fleet). In the sense of the DFA method, a taxi accident is essentially a random event. Factors such as culture, economics, climate, among others, do not generate memory effects in this system.

But if we compare these time series, we can see that some of them have cross-correlations. With the intention of checking their cross-correlations, we implement the DCCA method. The results of the DCCA analysis are found in Fig. 4 for cross-correlations between:

1. taxi driver age × time of driver licence (□),
2. taxi driver age × year of taxi manufacture (○),
3. taxi driver age × time of taxi repair (▽),
4. taxi driver age × cost of taxi repair (◇),
5. time of driver licence × year of taxi manufacture (×),
6. time of driver licence × time of taxi repair (⊗),
7. time of driver licence × cost of taxi repair (■),
8. year of taxi manufacture × time of taxi repair (▼),
9. year of taxi manufacture × cost of taxi repair (◆),
10. time of taxi repair × cost of taxi repair (∗).

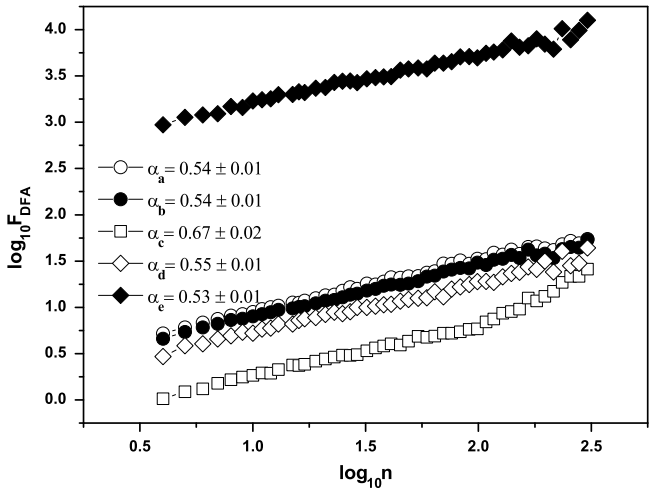

Fig. 3. DFA method applied to the case of taxi accidents for: (a) taxi driver age, (b) time of driver licence, (c) year of taxi manufacture, (d) time of taxi repair, and (e) the cost of taxi repair. For every case we show the angular coefficient for a linear adjust, e.g., the value of α_{DFA}.

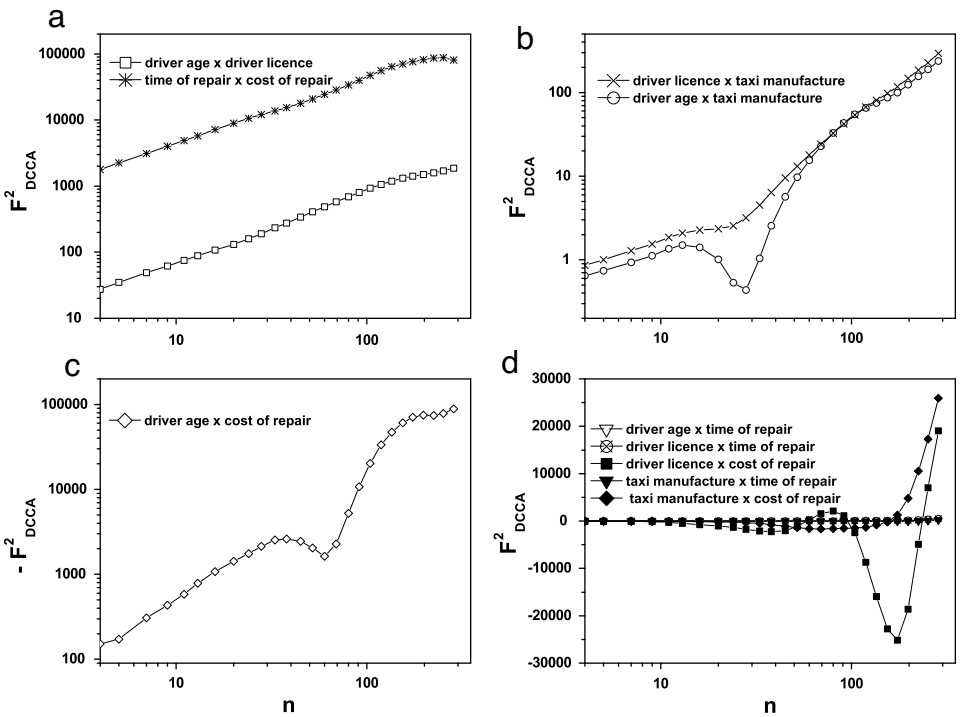

Fig. 4. DCCA analysis in case we have: (a) long-range cross-correlation, (b) positive short-range correlations, (c) negative short-range correlations, and (d) no cross-correlation between the time series.

If we look the behavior of the detrended covariance function in Fig. 4, we can divide the results into three distinct groups:

(i) where long-range cross-correlations are present ($F^2_{DCCA} \propto n^{2\lambda}$) (Fig. 4(a));
(ii) where there is a short-range cross-correlations, here with positive cross-correlation (see Fig. 4(b)) and negative cross-correlation (see Fig. 4(c));
(iii) where the detrended covariance function oscillates around zero, and we identify no cross-correlations between the time series (Fig. 4(d)).

Thus, specifically the cross-correlation between taxi driver age $\times$ time of driver licence ($\square$) and time of taxi repair $\times$ cost of taxi repair ($*$) (Fig. 4(a)), shows an evident long-range cross-correlation (group i) with $\lambda_{\square} = 0.52$ and $\lambda_{*} = 0.48$ (the power-law was measured by a linear fit with $\chi^2 > 0.98$ and $sd < 0.02$). However, by the DCCA method, we can see that there are other kinds of behavior in taxi accidents. For example, in Fig. 4(b) short-range cross-correlations appear between

Table 1
Matrix of cross-correlations between time series of taxi accidents recorded by the CEAT. The upper matrix represents the DCCA analysis, while the lower matrix represents the classical correlation coefficient (see Fig. 5). Here in this table LR = long-range cross-correlation, SR = short-range cross-correlation, and NC = no cross-correlation between the time series.

	Age	Lic.	Man.	Time	Cost
Age	♯♯♯	LR ($\lambda_{DCCA} = 0.52$)	SR (+)	NC	SR (−)
Lic.	0.85	♯♯♯	SR (+)	NC	NC
Man.	0.12	0.16	♯♯♯	NC	NC
Time	−0.01	0.01	0.10	♯♯♯	LR ($\lambda_{DCCA} = 0.48$)
Cost	−0.04	−0.02	0.07	0.44	♯♯♯

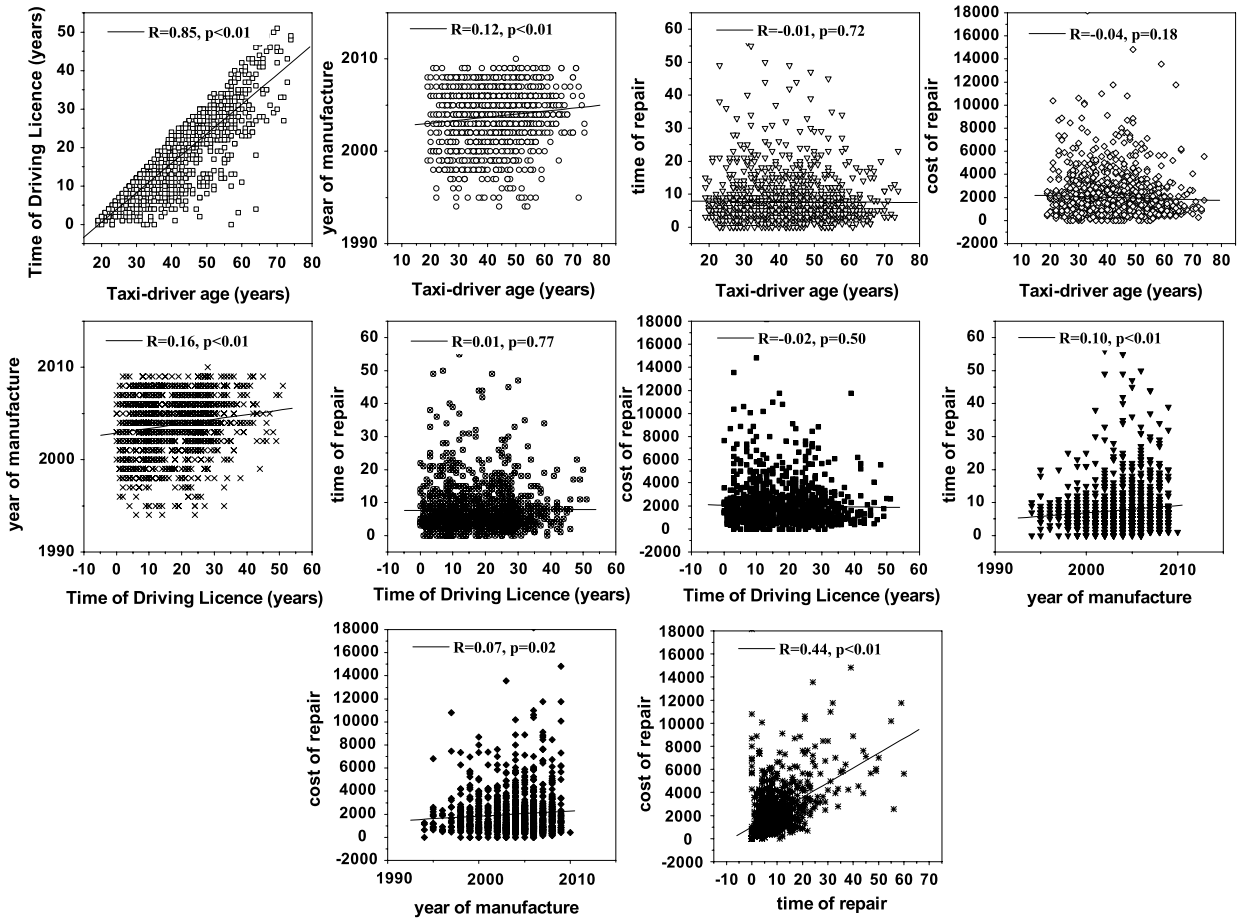

Fig. 5. This figure shows one variable against the other. Also shown in this figure is the R value and the *p*-value for the linear adjust (continuous line).

taxi driver age × year of taxi manufacture (○) and also time of driver licence × year of taxi manufacture (×). From this figure we can see that there is a slight tendency for taxi drivers with more experience to remain for a longer period of time with their taxi. Fig. 4(c) shows a short-range cross-correlation between taxi driver age × cost of taxi repair (◊), but we can prove by the DCCA method that this cross-correlation is negative. This figure presents a tendency for older taxi drivers to spend less money at the time of taxi repair, but as in Fig. 4(b), this cross-correlation is not long-range. Also, by the DCCA method, we can see that there is no cross-correlation between taxi driver age × time of taxi repair (▽), time of driver licence × time of taxi repair (⊗), time of driver licence × cost of taxi repair (■), year of taxi manufacture × time of taxi repair (▼), or year of taxi manufacture × cost of taxi repair (♦).

More succinctly, the DCCA cross-correlation analysis for taxi accidents can be summarized in Table 1.

In order to make a comparison between the methods, we plot one variable against the other (see Fig. 5 and Table 1 lower matrix). In this figure we propose a linear fit (continuous line) with the correlation coefficient values *R*, as well as the *p*-value. The results are equivalent to those obtained by the DCCA method (Fig. 4), but the DCCA is is more effective at identifying types of cross-correlation, especially because we can see F^2_{DCCA} for many time scales, and identify self-affinity in these time series. We can remember that the correlation coefficient detects only linear dependencies, while the DCCA

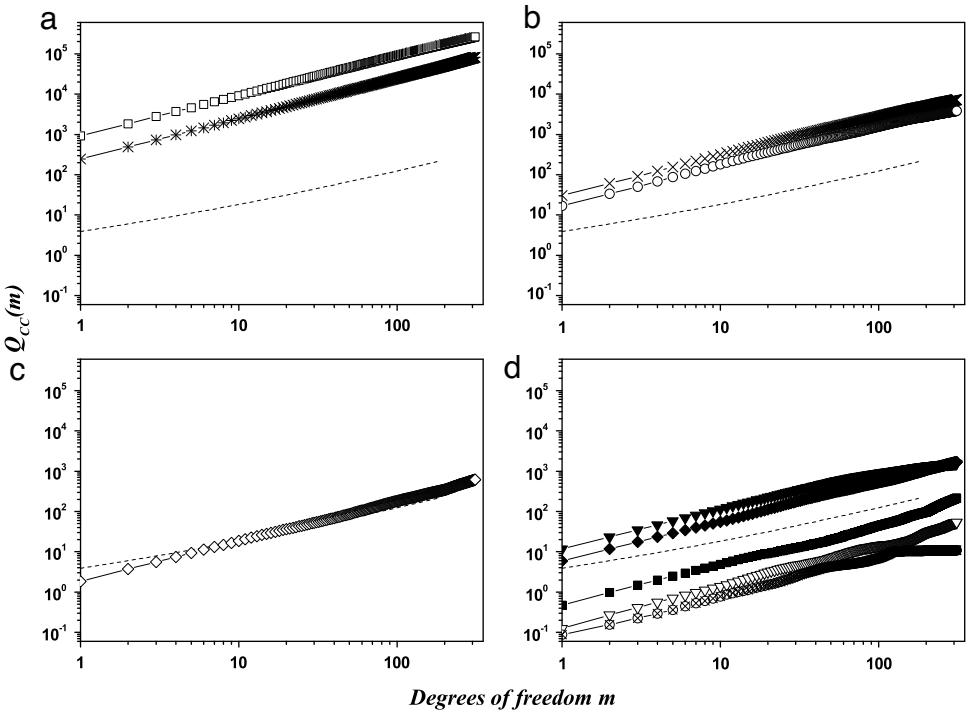

Fig. 6. The $Q_{cc}(m)$ qualitative test for cross-correlation between time series, where m is the number of degrees of freedom. The data are relative to the cross-correlations indicated in Fig. 4 (a)–(d). The dashed line denotes the critical values for the $\chi^2(m)$ distribution at the 5% level of significance.

method can remove polynomial trends. Later, we also test the cross-correlations analysis by the $Q_{CC}(m)$ function [34], for different choices of time lags (see Fig. 6). We can see that the qualitative test $Q_{CC}(m)$ [34] agrees with the quantitative DCCA cross-correlation method.

In conclusion, the DFA method indicates the presence of self-affinity in time series of taxi accidents, and the value of the α exponent is very close to 0.50 (uncorrelated time series), except for the year of taxi manufacture, where the value of α is 0.67 (this data can be identified with the renewal rate of the taxi fleet). In the sense of the DFA method, our analysis shows that a taxi accident is essentially a random event. On the other hand, with the DCCA method we can classify the cross-correlations into three types of groups, with long-range power-law cross-correlations, short-range cross-correlations, and one group where there are no cross-correlations. Finally, the data analysis using the DCCA method can lead to significant economic and social effects if, for example, we remember that the CEAT (companies) add value to the premium for a taxi, taking into account these variables. The DCCA method is shown here to be very consistent for data analysis and new observations can be found, as in this paper for taxi accidents.

Acknowledgements

We wish to thank the CEAT and FAPESB (Fundação de Amparo à Pesquisa do Estado da Bahia).

References

[1] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature (London) 356 (1992) 168.
[2] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, Z.D. Goldberger, S. Havlin, R.N. Mantegna, S.M. Ossadnik, C.-K. Peng, M. Simons, Physica A 205 (1994) 214.
[3] R.F. Voss, Phys. Rev. Lett. 68 (1992) 3805.
[4] G.F. Zebende, P.M.C. de Oliveira, T.J.P. Penna, Phys. Rev. E 57 (1998) 3311.
[5] R.L. Stratonovich, Topics in the Theory of Random Noise, Vol. 1, Gordon and Breach, New York, 1981.
[6] A. Witt, J. Kurths, A. Pikovsky, Phys. Rev. E 58 (1998) 1800.
[7] K. Hu, P.Ch. Ivanov, Z. Chen, P. Carpena, H.E. Stanley, Phys. Rev. E 65 (2002) 041107.
[8] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Phys. Rev. E 49 (1994) 1685.
[9] C.-K. Peng, S. Havlin, H.E. Stanley, A.L. Goldberger, Chaos 5 (1995) 82.
[10] K. Hu, P.Ch. Ivanov, Z. Chen, P. Carpena, H.E. Stanley, Phys. Rev. E 64 (2001) 011114.
[11] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E 51 (1995) 5084.
[12] M.A. Moret, G.F. Zebende, E. Nogueira Jr., M.G. Pereira, Phys. Rev. E 68 (2003) 041104.
[13] G.F. Zebende, M.G. Pereira, E. Nogueira Jr., M.A. Moret, Physica A 349 (2005) 452.
[14] G.F. Zebende, M.V.S. da Silva, A.C.P. Rosa Jr., A.S. Alves, J.C.O. de Jesus, M.A. Moret, Physica A 342 (2004) 322.
[15] L. Laloux, P. Cizeau, J. Bouchaud, M. Potters, Phys. Rev. Lett. 83 (1999) 1467.
[16] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, H.E. Stanley, Phys. Rev. Lett. 83 (1999) 1471.

[17] W.C. Jun, G. Oh, S. Kim, Phys. Rev. E 73 (2006) 066128.
[18] M. Campillo, A. Paul, Science 299 (2003) 547.
[19] P. Samuelsson, E.V. Sukhorukov, M. Büttiker, Phys. Rev. Lett. 91 (2003) 157002.
[20] W.-X. Zhou, Phys. Rev. E 77 (2008) 066211.
[21] A.L. Edwards, An Introduction to Linear Regression and Correlation, W. H. Freeman, San Francisco, 1976.
[22] M.R. Spiegel, Theory and Problems of Probability and Statistics, 2nd ed., McGraw-Hill, New York, 1992.
[23] T.S. Rao, M.B. Priestly, O. Lessi, Applications of Time Series Analysis in Astronomy and Meteorology, Chapman & Hall, Boca Raton, FL, USA, 1997.
[24] R.H. Shumway, D.S. Stoffer, Time Series Analysis and its Applications, with R Examples, 2nd ed., Springer-Verlag, New York, 2000.
[25] R.N. Mantegna, Eur. Phys. J. B 11 (1999) 193.
[26] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, T. Guhr, H.E. Stanley, Phys. Rev. E 66 (2002) 066126.
[27] L. Kullmann, J. Kertesz, K. Kaski, Phys. Rev. E 66 (2002) 026125.
[28] A. Cottet, W. Belzig, C. Bruder, Phys. Rev. Lett. 92 (2004) 206801.
[29] T. Mizunoa, H. Takayasu, M. Takayasu, Physica A 364 (2006) 336.
[30] B. Podobnik, D.F. Fu, H.E. Stanley, P.Ch. Ivanov, Eur. Phys. J. B 56 (2007) 47.
[31] K. Yamasaki, A. Gozolchiani, S. Havlin, Phys. Rev. Lett. 100 (2008) 228501.
[32] B. Podobnik, H.E. Stanley, Phys. Rev. Lett. 100 (2008) 084102.
[33] B. Podobnik, D. Horvatic, A.M. Petersena, H.E. Stanley, PNAS 106 (52) (2009) 22079.
[34] B. Podobnik, I. Grosse, D. Horvatic, S. Ilic, P.Ch. Ivanov, H.E. Stanley, Eur. Phys. J. B 71 (2009) 243.
[35] W.C. Jun, G. Oh, S. Kim, Phys. Rev. E 73 (2006) 066128.
[36] G.F. Zebende, A. Machado Filho, Physica A 388 (2009) 4863.