Missing Data Reconstruction in Acoustic Level Long Term Monitoring

CLAUDIO GUARNACCIA, JOSEPH QUARTIERI, CARMINE TEPEDINO
Department of Industrial Engineering
University of Salerno
Via Giovanni Paolo II 132, Fisciano (SA)
ITALY
cguarnaccia@unisa.it , quartieri@unisa.it , ctepedino@unisa.it

LJILJANA PETROVIC
Department of Mathematics and Statistics, Faculty of Economics
University of Belgrade
SERBIA
petrovl@ekof.bg.ac.rs

Abstract: - The acoustic level long term monitoring is a common practice in large urban areas, in which, according to the international regulation, the noise levels must be kept under certain thresholds. Sometimes, computational methods are used to predict acoustic level values in future periods. These models need to be calibrated on a continuous measurements dataset and, if some data are missing, the calibration can fail or can be "biased". In this paper, the problem of missing data reconstruction is approached by means of two techniques: a Time Series Analysis (TSA), based on the evaluation of trend and periodicity of the series, and a Regression (REGR) method, based on a modification of linear stochastic regression, will be presented and compared. The error analysis will show interesting features of both the models. In addition, the differences between a deterministic (TSA) and a stochastic imputation approach will be highlighted in terms of dataset mean and variance preservation.

Key-Words: - Acoustics, Imputation, Time Series Analysis, Regression Methods, Error Analysis

1 Introduction

One of the most important pollutants to be monitored and carefully controlled in urban areas is acoustical noise [1]. Its effects are largely documented in literature (see for instance [2-4]) and include both auditory and non-auditory possible damages to human health.

In order to monitor and predict the slope of noise produced by transportation infrastructures, that are the main sources together with industrial settlements, many models can be adopted, based on different approaches, such as experimental and software based analysis (for instance [5-16]), statistical methods and Poisson distributions (for instance [17-23]), etc..

The long term monitoring is a very helpful technique in order to control the daily and nightly levels and to check the variations over a large time range. The economical effort to install a noise level measurement station is not very high but the maintenance duties are very important, in order to have a continuous set of data. Sometimes, the station can have problems in collecting, recording or transmitting data, because of power fails, accidents, network problems, etc.. When one of these problems occurs, the dataset presents a "hole", a certain number of missing data.

The aim of this paper is to present and compare two models able to reconstruct the dataset and "fill the gaps" of missing data, according to Time Series Analysis (TSA) techniques or Regression (REGR) method. The performance of the models and the reliability of the reconstruction will be discussed in terms of comparison between actual data and predicted values. In addition, an error analysis will be reported and discussed in the last section, looking at the difference between observed and simulated values.

2 Models Presentation

The models presented in this paper are based on two methods:

- Time Series Analysis model (TSA)
- Regression method (REGR)

Both of them are largely adopted in many applications and can be considered for missing data reconstruction issues.

2.1 Time Series Analysis model
The TSA model is a mathematical model able to reproduce the behavior of a certain time series, by estimating its trend and its periodic pattern. These two elements do not include the random component, that is non-deterministic, and that cannot be predicted in advance.

The TSA models are used in many disciplines [24-33] and, according to how the trend, the seasonality and the random term are composed, can be multiplicative, additive or mixed. The model adopted in this paper is a mixed one (multiplicative between trend and seasonality, with the adding of the random term), and it is based on the following formula:

$$Y_t = T_t S_t + r_t \quad , \qquad (1)$$

where Y_t is the Time Series, T_t is the trend, S_t is the periodic component and r_t is the random term.

The random component can be evaluated on a calibration dataset as the difference between the observed data and the "punctual forecast", i.e. the product of trend and seasonality. In this way, a distribution of the random term is obtained and the mean can be used in the final forecast (F_t) formula:

$$F_{t,TSA} = T_t \bar{S}_i + m_e \quad , \qquad (2)$$

where $\bar{S}_i$ is the seasonal coefficient and m_e is the mean of the random component.

Of course one should expect that, if the periodicity of the data is completely included in the seasonal coefficients and the trend is properly evaluated, the mean of the random term (also called "noise") should be zero, i.e. the error distribution is normal and centred in the zero value. If this does not happen, the mean can be added to improve the predictive model [27, 28]. The presence of a further periodicity in the data can be also evidenced evaluating the autocorrelation of the error (difference between observed and predicted values).

In this paper, the calibration of the TSA model has been done on different datasets, according to the validation test ongoing. When reconstructing a 20 data interval, the calibration has been done on the entire dataset, minus the 20 missing data. On the contrary, when reconstructing 60 data, 20 at the beginning of the time series, 20 in the middle and 20 randomly cut, the calibration has been done on the entire dataset minus the 60 missing data. The point is that the model parameters are never evaluated on the data that need to be reconstructed. In fact, the best applications for TSA are related to the prediction of data not used in the calibration phase, as reported in [27-33].

2.2 Regression method
Missing (or incomplete) data are a part of almost all research, and one has to decide how to deal with it each time. When the missing data comprise only a small fraction of all cases (say, five percent or less) then case deletion may be a perfectly reasonable solution to the missing data problem. In other cases, the imputation method is used in statistical practice to fill in missing data with plausible values.

The methods usually used in time series analysis for filling gaps are methods which replace the missing data with series mean, or with mean (or median) of nearby points or method of linear interpolation.

There are a number of alternative ways of dealing with missing data. A very useful literature to understand both the theoretical and practical implications of the different methods to deal with missing data is, for example, [34] and [35].

Any discussion of missing data must begin with the question of why data are missing. The reasons for missing data plays an important role in how those data will be treated. The method discussed here requires that the data are "Missing At Random" (MAR) – i.e. not related to the missing values. Precisely, data are missing at random if the probability of missing data on a variable is not a function of its own value after controlling for other variables in the design. Despite its name, MAR does not suggest that the missing data values are a simple random sample of all data values. To estimate missing data, in this paper a modification of linear stochastic regression is considered. More precisely, the following procedure is applied: first, the missing values are replaced with the mean of all available data, and (then) regression estimation is improved by adding a random normal variable to each estimate (i.e. an error component is added to each observation):

$$F_{t,REGR} = \bar{y} + u_t \quad , \qquad (3)$$

where $\bar{y}$ is the mean of the "calibration" data and u_t is the random term, drawn from a normal distribution with the expected value 0 and the standard deviation equal to the square root of the mean squared error term of the regression.

This method restores some lost variability of data.

ISBN: 978-1-61804-315-3

3 Analysis and results

In order to evaluate the performances of the two methods, a long term noise measurements dataset has been considered. It is a set of noise level measurements recorded in Messina, Italy, all day long. In this paper, the authors consider the day time measurements related to the site of Viale Boccetta, from the 11th of May 2007 to the 26th of March 2008, that are 321 data. From this dataset, the last 21 data have been separated from the main interval, to be used in a validation phase, that will be reported in a further paper. Thus, 300 of data are available for the reconstruction analysis presented in this section.

Three "cuts" have been done in the 300 data interval, all of them of 20 data. The first cut has been done at the beginning of the dataset (data 1-20), the second in the middle (data 150-169) and the third randomly all over the dataset. This has been done to check if the reconstruction techniques are sensible to the position of the missing data in the dataset.

In addition, a comparison on 60 missing data has been done, recalibrating the models on the remaining 240 data. This analysis will show the performances of the models in reconstructing a quite large number of missing data (20% of the total).

3.1 Time history comparison

The first analysis that can be performed is based on the plot of the real data that have been cut, together with the "reconstructed" data.

In Fig. 1 the three cut intervals are considered and the two models results are plotted with the data removed from the dataset.

It is easy to notice that the TSA curve is closer to real data with respect to REGR. This is probably due to the fact that TSA model is "trained" on the entire dataset and the parameter evaluation has been performed considering all the data. In addition, the REGR method is non deterministic and it is strongly influenced by the stochastic approach.

3.2 Reconstruction error evaluation

A quantitative analysis can be done evaluating the error e_t of the models, according to the following formula:

$$e_t = A_t - F_t \quad , \qquad (4)$$

where A_t is the "actual" value and F_t the forecast.

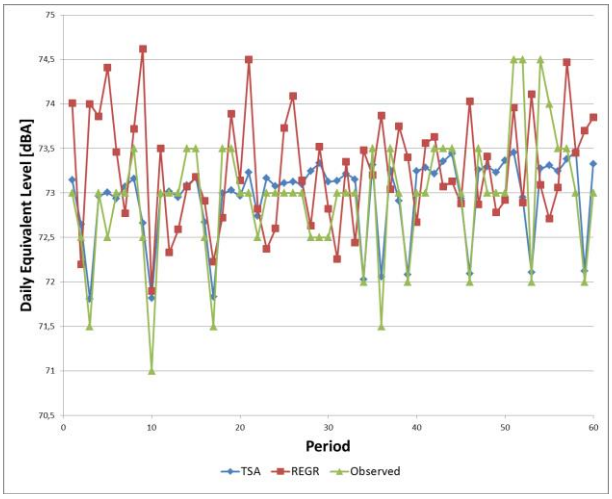

Fig. 1: Comparison between observed (green line) and reconstructed data with TSA model (blue line) and REGR method (red line). The *x* axis is the union of the 1-20, 150-169 and random cut data intervals.

3.2.1 Analysis on 20 data cut intervals

Considering the 20 data cut intervals, the two methods have been calibrated and applied to reconstruct the missing data. Let us underline that models are calibrated on 280 data, that are the 300 complete dataset minus 20 cut data (each time cut in different positions).

The mean errors and other statistics are reported in Tab. 1 for the three "cut intervals" in the data.

Tab. 1: Summary of error statistics for the 20 data cut intervals.

		Mean [dBA]	Std.dev [dBA]	Median [dBA]	Skew	Kurt
1-20	REGR	-0,10	0,79	0,03	-0,47	-0,67
	TSA	-0,01	0,33	0,00	-0,20	-0,31
150-169	REGR	0,24	0,99	0,04	0,52	-0,88
	TSA	0,19	0,55	0,08	1,13	0,26
Random	REGR	-0,32	0,78	-0,48	0,26	-0,10
	TSA	-0,21	0,29	-0,16	-0,61	-0,23

Let us remind that a negative error occurs when the model prediction is higher than the actual value (overestimation).

Results in Tab. 1 show that, the REGR method has a higher mean error with respect to TSA, in all the intervals. Also the standard deviation of the error is higher.

In Figg. 2-4, the plots of the errors in the three cut intervals are reported. Of course, the more the curve approaches zero, the best the model reconstruction performance is.

It can be noticed that the error curves are slightly affected by the "position" of the cut interval. In fact, in the 1-20 cut interval (Fig. 2), the errors are lower than the other cases.

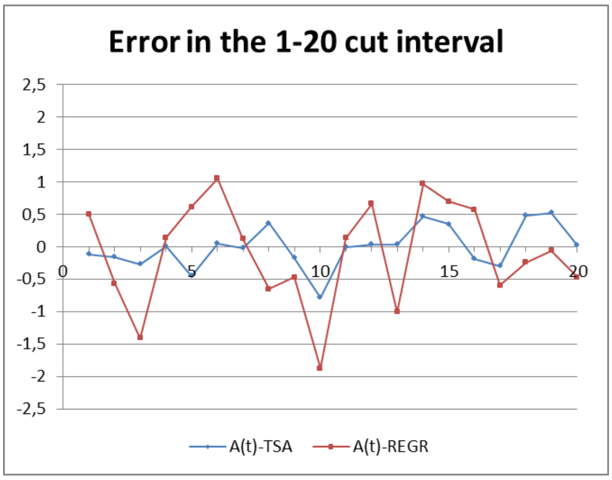

Fig. 2: Errors [dBA] of Time Series Analysis (TSA) model (blue line) and Regression (REGR) method (red line) in the 1-20 data reconstruction (at the beginning of the series).

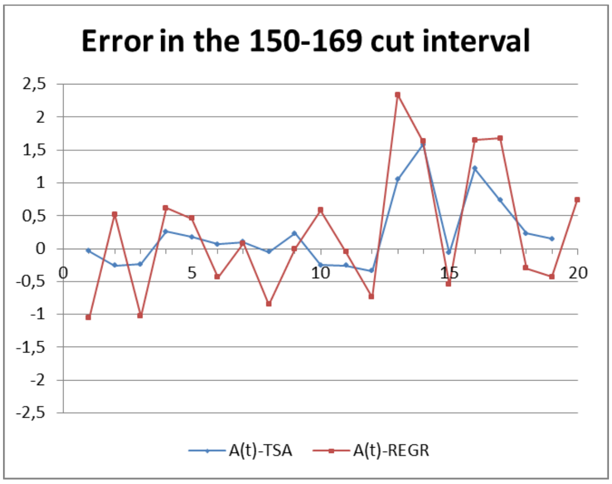

Fig. 3: Errors [dBA] of Time Series Analysis (TSA) model (blue line) and Regression (REGR) method (red line) in the 150-169 data reconstruction (in the middle of the series).

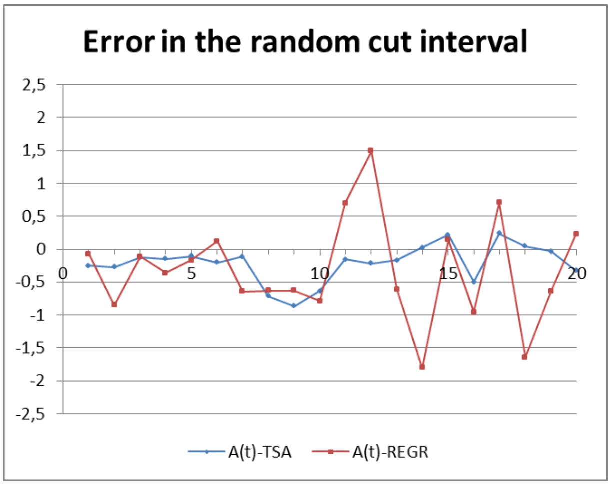

Fig. 4: Errors [dBA] of Time Series Analysis (TSA) model (blue line) and Regression (REGR) method (red line) in the random cut data reconstruction.

In 150-169 range (Fig. 3), a quite evident underestimation is present in the second part of the interval, for both models.

In the random cut interval (Fig. 4), there is a quite variable error, in particular for REGR model. This is probably due to the random component added in the REGR technique.

3.2.2 Analysis on 60 data cut interval

In this subsection, the analysis on reconstruction of 60 data, cut at the beginning (1-20), in the middle (150-169) and randomly, is presented. The models are calibrated on the resulting dataset of 240 data (300 data minus the missing intervals).

In Fig. 5, the plot of the errors in the reconstruction of all the 60 cut data is reported for both models. Let us underline that this is not the composition of the previous three plots (Figg. 2-4), because in this case, both the models are calibrated on 240 data, that are the 300 complete dataset minus 60 cut data. In addition, since REGR is a stochastic model, every run of the model, even on the same dataset, give different results, because of the randomly generated component.

The statistics of the error are reported in Table 2. A general overestimation is evidenced for both model and the performances of REGR are consistent with that in the 20 missing data cases.

In particular, it can be observed that the REGR method approaches the highest mean and standard deviation values of Table 1, while TSA tends to the lowest mean value and an average standard deviation.

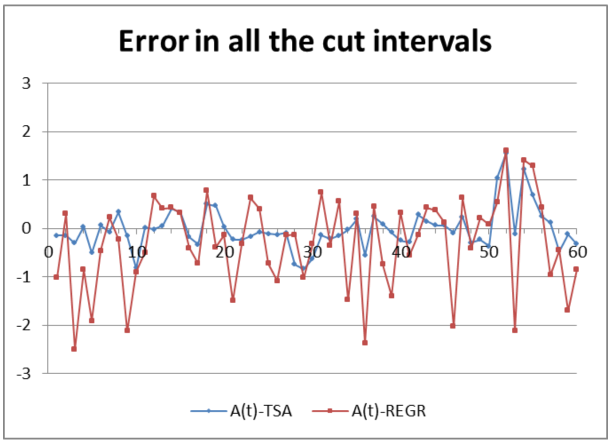

Fig. 5: Errors [dBA] of Time Series Analysis (TSA) model (blue line) and Regression (REGR) method (red line) in the complete 60 cut data reconstruction (1-20, 150-169 and random cut).

Tab. 2: Summary of error statistics for the 60 cut data interval.

		Mean [dBA]	Std.dev [dBA]	Median [dBA]	Skew	Kurt
60 clipped data	REGR	-0,32	0,95	-0,27	-0,42	-0,39
	TSA	-0,02	0,43	-0,10	1,18	2,72

3.3 Dataset reconstruction performances

A further analysis can be performed looking at the entire dataset. The complete observed time series can be compared with the datasets reconstructed with TSA and REGR techniques. This comparison will furnish a quantitative idea of how the different imputation techniques affect the mean value and the distribution variance of the dataset.

In this subsection, the models calibrated on 240 data (entire dataset, 300 data, minus the 60 missing data as in subsection 3.2.2) are considered. In particular, the datasets considered here are composed by the 240 observed data plus the 60 reconstructed data by means of the two models.

Since TSA is a deterministic model, because once the parameters are evaluated the forecast in a given time *t* is always the same, the mean is preserved. On the contrary, since REGR is a stochastic imputation technique, i.e. the forecast on a given time *t* changes each time the model runs, the distribution (variance) is preserved. [36]

Thus, one should expect that the mean of the data obtained with the TSA model is very close to the mean of the observed data. The same is expected for the variance obtained with the REGR imputation method. These considerations, even if with very slight differences, are confirmed by the results reported in Table 3.

Tab. 3: Summary of observed and simulated dataset statistics.

	Mean [dBA]	Std.dev [dBA]	Min [dBA]	Max [dBA]
Observed	73,07	0,66	70,5	75
REGR	73,13	0,64	70,5	75
TSA	73,07	0,61	70,5	75

4 Conclusions

In this paper, the problem of missing data in long term acoustic level monitoring has been considered.

In particular, the reconstruction of missing data by means of Time Series Analysis (TSA) model and Regression method has been pursued.

The TSA reconstruction is based on the evaluation of trend and periodicity of the series, and completed with the adding of the mean of the error in the calibration phase.

The REGR imputation is based on a modification of linear stochastic regression, in particular on the calculation of the mean of the available calibration data and the adding of a random stochastic component.

The analysis of the error, i.e. the difference between observed and reconstructed values, evidenced that TSA has better reconstruction performances, both in terms of mean and standard deviation of the error. This analysis has been performed on three different intervals, all of them of 20 data, the first at the beginning of the dataset (data 1-20), the second in the middle (data 150-169), the third with a random position of the 20 cut data. The differences between the position of the cut interval are very small and the general behavior of the models is preserved.

An additional analysis has been performed on a 60 cut data interval (1-20, 150-169 and random cut), to check any variation of the models reconstruction performances when a high percentage of the data is missing (20%). In this case, it has been observed that the REGR method approaches the highest mean and standard deviation values of the 20 cut data cases, while TSA tends to the lowest mean value and an average standard deviation.

Finally, an analysis on the 300 data (240 observed plus 60 reconstructed with the two models) has been implemented, confirming what reported in literature, i.e. that deterministic methods (such as TSA) preserve the mean of the data, while stochastic methods (such as REGR) preserve the distribution variance.

Future steps of this work can be the inclusion of other recorded variables for estimation the daily acoustic level as main variable. This will lead to better results of the REGR model.

References:

[1] WHO, World Health Organization, *Guidelines for community noise*, in: Berlund B, Lindvall T, Schewela DH, editors. Geneva: World Health Organization; 1999. p. v–xviii, 19–20.

[2] Kryter KD, *The Effects of Noise on Man*, 2nd edition, Orlando, FL, Academic Press, 1985.

[3] Ouis D., Annoyance from road traffic noise: a review, *Journal of Environmental Psychology*, 21, 2001, 101-120

[4] Guarnaccia C., Mastorakis N. E., Quartieri J., Noise Sources Analysis in a Wood Manufacturing Company, *International Journal of Mechanics*, Issue 2, Vol. 7, pp 37-44 (2013), ISSN: 1998-4448.

[5] Quartieri J., Mastorakis N. E., Iannone G., Guarnaccia C., D'Ambrosio S., Troisi A., Lenza T.L.L, *A Review of Traffic Noise Predictive Models*, Proc. of the 5th WSEAS International Conference on "Applied and Theoretical Mechanics" (MECHANICS'09), Puerto de la Cruz, Tenerife, Spain, 14-16 December 2009, pp. 72-80.

[6] Guarnaccia C., Lenza T.LL., Mastorakis N.E., Quartieri J., A Comparison between Traffic Noise Experimental Data and Predictive Models Results, *International Journal of Mechanics*, Issue 4, Vol. 5, pp. 379-386 (2011), ISSN: 1998-4448.

[7] Guarnaccia C., Analysis of Traffic Noise in a Road Intersection Configuration, *WSEAS Transactions on Systems*, Issue 8, Volume 9, (2010), pp.865-874, ISSN: 1109-2777.

[8] Quartieri J., Mastorakis N. E., Guarnaccia C., Troisi A., D'Ambrosio S., Iannone G., Traffic Noise Impact in Road Intersections, *International Journal of Energy and Environment*, Issue 1, Vol. 4 (2010), pp. 1-8.

[9] Guarnaccia C., *Acoustical Noise Analysis in Road Intersections: a Case Study*, Proceedings of the 11th WSEAS International Conference on "Acoustics & Music: Theory & Applications" (AMTA '10), Iasi, Romania, 13-15 June 2010, pp. 208-215.

[10] Quartieri J., Troisi A., Guarnaccia C., Lenza TLL, D'Agostino P., D'Ambrosio S., Iannone G., An Acoustical Study of High Speed Train Transits, *WSEAS Transactions on Systems*, Issue 4, Vol.8, pp. 481-490 (2009).

[11] Quartieri J., Troisi A., Guarnaccia C., Lenza TLL, D'Agostino P., D'Ambrosio S., Iannone G., Application of a Predictive Acoustical Software for Modelling Low Speed Train Noise in an Urban Environment, *WSEAS Transactions on Systems*, Issue 6, Vol.8, pp. 673-682 (2009).

[12] Quartieri J., Troisi A., Guarnaccia C., Lenza TLL, D'Agostino P., D'Ambrosio S., Iannone G., *Analysis of Noise Emissions by Train in Proximity of a Railway Station*, Proceedings of the 10th International Conference on "Acoustics & Music: Theory & Applications" (AMTA '09), Prague (Rep.Ceca), 23-25 March 2009, pp: 100-107.

[13] Quartieri J., Troisi A., Guarnaccia C., Lenza TLL, D'Agostino P., D'Ambrosio S., Iannone G., *An Italian High Speed Train Noise Analysis in an Open Country Environment*, Proceedings of the 10th International Conference on "Acoustics & Music: Theory & Applications" (AMTA '09), Prague (Rep.Ceca), 23-25 March 2009, pp: 92-99.

[14] Quartieri J., Mastorakis N. E., Guarnaccia C., Troisi A., D'Ambrosio S., Iannone G., *Road Intersections Noise Impact on Urban Environment Quality*, Proceedings of the 5th WSEAS International Conference on "Applied and Theoretical Mechanics" (MECHANICS '09), Puerto de la Cruz, Tenerife, Spain, 14-16 December 2009, pp. 162-171.

[15] Guarnaccia C., Quartieri J., Ruggiero A., Acoustical Noise Study of a Factory: Indoor and Outdoor Simulations Integration Procedure, *International Journal of Mechanics*, Vol. 8, pp. 298-306 (2014), ISSN 1998-4448.

[16] Guarnaccia C., Mastorakis N.E., Quartieri J., Wind Turbine Noise: Theoretical and Experimental Study, *International Journal of Mechanics*, Issue 3, Vol.5, pp. 129-137 (2011).

[17] Iannone G., Guarnaccia C., Quartieri J., Speed Distribution Influence in Road Traffic Noise Prediction, *Environmental Engineering And Management Journal*, Vol. 12, Issue 3, 2013, pp. 493-501.

[18] Quartieri J., Iannone G., Guarnaccia C., *On the Improvement of Statistical Traffic Noise Prediction Tools*, Proceedings of the 11th WSEAS International Conference on "Acoustics & Music: Theory & Applications" (AMTA '10), Iasi, Romania, 13-15 June 2010, pp. 201-207.

[19] Iannone G., Guarnaccia C., Quartieri J., *Noise Fundamental Diagram deduced by Traffic Dynamics*, in "Recent Researches in Geography, Geology, Energy, Environment and Biomedicine", Proc. of the 4th WSEAS Int. Conf. on Engineering Mechanics, Structures, Engineering Geology (EMESEG '11), Corfù Island, Greece, July 14-16, 2011, pp. 501-507.

[20] Quartieri J., Mastorakis N.E., Guarnaccia C., Iannone G., *Cellular Automata Application to Traffic Noise Control*, Proc. of the 12th Int. Conf. on "Automatic Control, Modelling & Simulation" (ACMOS '10), Catania (Italy), 29-31 May 2010, pp. 299-304.

[21] C. Guarnaccia, J. Quartieri, C. Tepedino, E. R. Rodrigues, An analysis of airport noise data using a non-homogeneous Poisson model with a change-point, *Applied Acoustics*, Vol. 91, pp. 33-39, 2015, doi:10.1016/j.apacoust.2014.12.002

[22] C. Guarnaccia, J. Quartieri, J. M. Barrios, E. R. Rodrigues, Modelling Environmental Noise Exceedances Using non-Homogenous Poisson Processes, *Journal of the Acoustical Society of*

America, 136, pp. 1631-1639 (2014) http://dx.doi.org/10.1121/1.4895662
[23] Guarnaccia C., Advanced Tools for Traffic Noise Modelling and Prediction, *WSEAS Transactions on Systems*, Issue 2, Vol.12, 2013, pp. 121-130.
[24] Brockwell P.J, Davis R.A., *Introduction to Time Series and Forecasting*, second edition, Springer Texts in Statistics, 2002.
[25] Pope C.A., Dockery D.W., Spengler J.D., and Raizenne M.E., Respiratory Health and PM10 Pollution: A Daily Time Series Analysis, *American Review of Respiratory Disease*, Vol. 144, No. 3_pt_1, (1991) pp. 668-674.
[26] Milanato D., *Demand Planning. Processi, metodologie e modelli matematici per la gestione della domanda commerciale*, Springer, Milano, 2008, in Italian.
[27] Guarnaccia C., Quartieri J., Mastorakis N.E., Tepedino C., Development and Application of a Time Series Predictive Model to Acoustical Noise Levels, *WSEAS Transactions on Systems*, Vol. 13, (2014) pp. 745-756, ISSN / E-ISSN: 1109-2777 / 2224-2678
[28] Guarnaccia C., Quartieri J., Rodrigues E.R., Tepedino C., Acoustical Noise Analysis and Prediction by means of Multiple Seasonality Time Series Model, *International Journal of Mathematical Models and Methods in Applied Sciences*, Vol. 8, (2014) pp 384-393, ISSN: 1998-0140.
[29] Guarnaccia C., Cerón Bretón J.G., Quartieri J., Tepedino C., Cerón Bretón R.M., An Application of Time Series Analysis for Forecasting and Control of Carbon Monoxide Concentrations, *International Journal of Mathematical Models and Methods in Applied Sciences*, Vol. 8, (2014) pp 505-515, ISSN: 1998-0140
[30] Tepedino C., Guarnaccia C., Iliev S., Popova S. and Quartieri J., A Forecasting Model Based on Time Series Analysis Applied to Electrical Energy Consumption, *International Journal of Mathematical Models and Methods in Applied Sciences*, accepted and in press, 2015.
[31] Guarnaccia C., Quartieri J., Mastorakis N. E. and Tepedino C., *Acoustic Noise Levels Predictive Model Based on Time Series Analysis*, in Latest Trends in Circuits, Systems, Signal Processing and Automatic Control, proceedings of the 2nd Int. Conf. on Acoustics, Speech and Audio Processing (ASAP '14), Salerno, Italy, June 3-5, 2014, ISSN: 1790-5117, ISBN: 978-960-474-374-2, pp. 140-147.
[32] Guarnaccia C., Quartieri J., Rodrigues E. R. and Tepedino C., T*ime Series Model Application to Multiple Seasonality Acoustical Noise Levels Data Set*, in Latest Trends in Circuits, Systems, Signal Processing and Automatic Control, proceedings of the 2nd Int. Conf. on Acoustics, Speech and Audio Processing (ASAP '14), Salerno, Italy, June 3-5, 2014 , ISSN: 1790-5117, ISBN: 978-960-474-374-2, pp. 171-180.
[33] Guarnaccia C., Quartieri J., Cerón Bretón J. G., Tepedino C., Cerón Bretón R. M., *Time Series Predictive Model Application to Air Pollution Assessment*, in “Latest Trends on Systems”, Proc. of the 18th Int. Conf. on Circuits, Systems, Communications and Computers (CSCC'14), Santorini, Greece, 17-21 July 2014, pp. 499-505.
[34] Schafer J. L., *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall, 1997.
[35] Little R.J.A. and Rubin D.B., Statistical Analysis with Missing Data. New York: John Wiley & Sons, 1987.
[36] Chen J. N. K. Rao and R. R. Sitter, Efficient Random Imputation for Missing Data in Complex Surveys, Statistica Sinica 10 (2000), 1153-1169.