A MULTIPLE REGRESSION APPROACH TO FORECASTING PM10 CONCENTRATION IN THE CITY OF PADUA, ITALY

Maria Sansone*, Massimo Bressan**, Denise Pernigotti*, Andrea Rossa*,
Massimo Ferrario* and Alessandro Benassi*
*ARPAV-Meteorological Center of Teolo, Teolo (PD), Italy; **ARPAV-Department of Padua, Padua, Italy

Abstract
Particulate matter (PM_{10}) concentration at ground level is strongly affected by meteorological conditions. This study presents a multiple regression approach to daily average PM_{10} concentration using log-normal variables transformation in order to better understand which factors are more appropriate for PM_{10} forecasting. The meteorological factors used in this linear regression are daily mean variables of wind velocity, rain accumulation, mixing height, thermal inversion index. The estimation of the multiple regression coefficients were done on the basis of a data set monitored in the urban area of Padua, Italy during the period October 2001 – September 2004. Determination coefficient R^2 , used to test the fitness of the regression, was 0.75 when evaluated on the same dataset and 0.72 when evaluated on the winter period October 2004 – September 2005. A forecast test using two day before PM_{10} concentration, two day before meteorology and one day before meteorology is done, to test the forecast reliability of the regression approach. In this case the determination coefficient was on the analysis dataset and in the verification period.

Key words: PM_{10}, meteorology, multiple regression, forecast

1. INTRODUCTION

High concentrations of particulate matter PM_{10} in urban areas have a serious impact on human health. Good understanding and reliable forecasting of PM10 concentrations allow well-timed population information about urban air quality and effective policy decision making such as local traffic management (circulation restrictions) and/or large scale regional reduction programmes. The city of Padua is located in a lowland (Pianura Padana) where is likely to occur more than 150 excedences per year of the threshold of 50 $\mu g/m^3$ and more than 15 exceedences per year of 100 $\mu g/m^3$ PM10 daily concentrations.
On the other hand deterministic dispersion models are not yet fully capable to couple with the order of magnitude of such strong and short peak events of pollution neither in forecasting nor in analysis.
The idea here is to prove that such high peak events can be explained with the peculiar state of the planetary boundary layer (PBL) in a lowland where effects of stagnation are highly enhanced especially during winter season. Once we understand which PBL parameters are the most important in controlling PM10 concentrations at ground level we can also try to apply this statistical approach to routinely daily forecasting.

2. DESCRIPTION OF THE DATA SET

The period selected for analysis is from October 2001 to September 2004. PM_{10} concentrations considered in the regression analysis was the average of daily concentrations measured at two monitoring stations placed in the urban area of Padua: *Arcella* (traffic hot spot) and *Mandria* (urban background).
The meteorological data, averaged daily, were measured at the closest CMT (Meteorological Center of Teolo) station of *Legnaro* which is about 10 km SE of the city Center of Padua.
The valid data of this period are 981 out of 1096 day considered (89.5%). When we considered the day-2 regression the valid data decreased to 956 (87.2%).

Since October 2003 Meteorological Center of Teolo gives a daily forecast bulletin for PM_{10}, based on subjective evaluation of the weather forecast. By using this experience in PM_{10} forecasting and after a quick preliminary correlation coefficient analysis some PBL variables have been selected: wind velocity, rain accumulation, mixing height, thermal inversion index .

The mixing height (Hmix) is the top of the PBL for which the most accepted definition is by *Stull (1998*): "*the part of troposphere that is directly influenced by the presence of the earth's surface, and respond to surface forcings with a timescale of about one hour or less".* Mixing height is calculated with the method of the energy balance proposed by various authors (see bibliography in *Scire, 2000)*, also used in US-EPA meteorological preprocessors for dispersion modeling like METRO, AIRMET, CALMET. During very stable condition this algorithm accept a user defined minimum value without discrimination between different stability strength.

For this reason a thermal inversion index (*Stanford)* was also calculated, which is supposed to be proportional to the inversion capping effects as defined by the following formula:

$$I = \frac{(\Delta\theta)^2}{3 + (z+1)\cdot\Delta z}$$

where $\Delta\theta$ is the potential temperature difference between top and bottom of the inversion (K), z is the height of the inversion bottom (hm) and Δz is the depth of the inversion (hm). The vertical profile of potential temperature was

calculated with the interpolation of TEMP data measured in Udine (16044), Milano Linate (16080) and S. Pietro Capofiume (16144) plus surface temperature data from meteorological station placed in Legnaro (Padua).

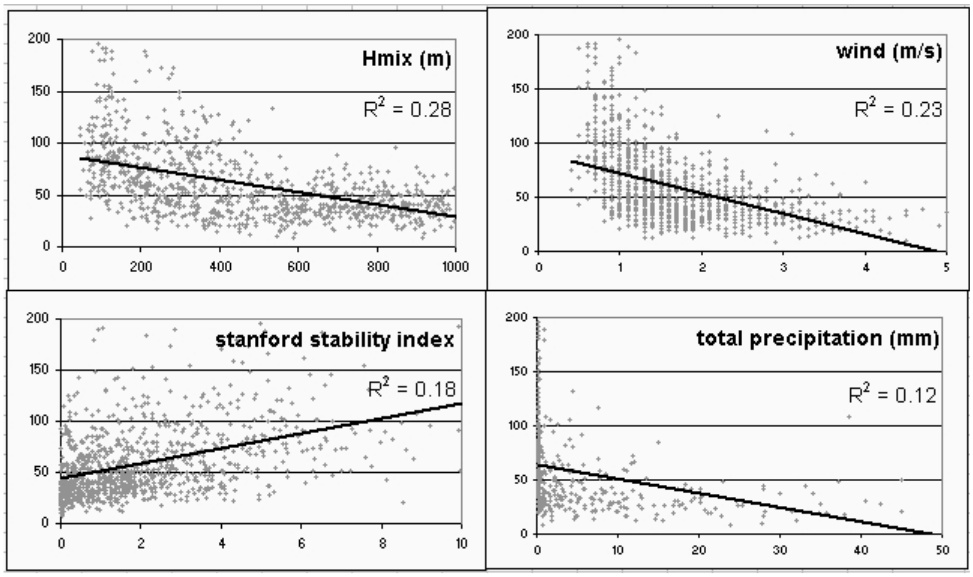

Figure 1: scatter plot of PM_{10} ($\mu g/m^3$, y axis) vs mixing height, wind velocity, stanford stability index and total precipitation, all daily averaged

Figure 1 illustrates how above mentioned variables correlate to PM_{10} concentration. It is important to notice that only Stanford index is positively correlated. The determination coefficient R^2 for PM_{10} autocorrelation is high: 0.63 day-1, 0.35 day-2, 0.20 day-3 and 0.13 day-4.

3. MULTIPLE LOG-NORMAL LINEAR REGRESSION MODEL

In general terms, multiple regression procedures will estimate a linear equation of the form:

$$Y= a+b_1*X_1 + b_2*X_2 +...+ b_p*X_p$$

Note that in this equation, the regression coefficients (or b coefficients) represent the contributions of each independent variable to the prediction of the dependent variable. Another way to express this fact is to say that, for example, variable X_1 is correlated with the Y variable, after controlling for all other independent variables.
The degree to which two or more predictors (independent or X variables) are related to the dependent (Y) variable is expressed in the correlation coefficient R, which is the square root of R-square. In multiple regression, R can assume values between 0 and 1. To interpret the direction of the relationship between variables, one looks at the signs (plus or minus) of the regression or B coefficients. If a B coefficient is positive, then the relationship of this variable with the dependent variable is positive; if the B coefficient is negative then the relationship is negative. Of course, if the B coefficient is equal to 0 then there is no relationship between the variables. In this application the independence of the variables is reasonable but probably not completely true.
In the linear regression here applied the assumption of constant emissions is made.
The weights and the statistics for a linear regression are reported in the following Table, it can be seen that in this case the intercept a is not null.

Statistics on regression vs day-0 PM10			regression coefficients					
				day-0				day-1
R^2	explaned variance	RMSE	a	wind velocity	hmix	standford	total precip.	PM10
0.75	29.57	16.95	43.25	-8.73	-0.02	1.68	-0.41	0.63

Table 1: statistical scores and coefficients for regression with original data variables

3.2. Log-normal regression

Another assumption is that variables have a log-normal distribution which is a rigorous hypothesis for PM_{10} concentrations (*Cacciamani, 2001*) as confirmed by the Kolmogorov-Smirnov test for normality (d=0.03192, p>0.20) shown in Figure 2, but result just as a good approximation hypothesis for meteorological variables such as, mixing height and Stanford index.

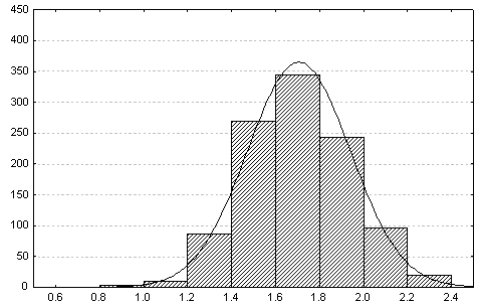

Figure 2: hystogram of the K-S test for log-normal PM_{10} concentration distribution

Normalized data for mean and standard deviation of all variables are considered in the regression model. Result showed that regression coefficients ("weights") are directly proportional to the importance of associated physical process of PM_{10} dispersion in the low atmosphere.

The following table shows the average and standard deviation of logarithm of variables used for normalization of input data in the regression model.

	wind velocity	hmix	standford	total precip.	PM10avg
average	0.43	5.95	0.04	-1.36	3.93
standard deviation	0.46	0.75	1.56	1.76	0.55

Table 2: descriptive statistics of logaritms of data, used for renormalization

The linear regression was calculated with the MATLAB function "regress" (see bibliography in MATLAB Manual) and best results are showed in Table 3. A "step in step out" test with Stanford index was done showing how excluding this variable the fitting of the regression resulted just a little worst (by a factor of 10^{-2}). Anyhow we believed that this variables should be important in caching PM_{10} peaks and therefore we decided to take it into the regression just for day-0.

Statistics on regression vs day-0 PM10			regression coefficients								
			day-0				day-1				day-2
R^2	explaned variance	RMSE	wind velocity	hmix	standford	total precip.	PM_{10}	wind velocity	hmix	total precip.	PM_{10}
0.75	0.87	0.50	-0.23	-0.19	0.08	-0.18	0.55				
0.68	0.83	0.56	-0.23	-0.28	0.07	-0.15		-0.10	-0.02	-0.20	0.30

Table 3: statistical scores and coefficients for the multilinear regression with log-normal variables

The most important meteorological parameter at day-0 result to be the mixing height, which at day-1 is much less important than total precipitation and wind velocity. On the contrary total precipitation is even more important at day-1, when it becomes the most important meteorological parameter.

From a statistical point of view the determination coefficient R^2 using just PM_{10} day-2 is comparable with the autocorrelation of PM_{10} day-1 (R^2=0.63), which could mean that, in the short-time, meteorology completely explains PM_{10} concentrations and therefore the initial assumption of constant emissions is a good approximation.

3.2. Verification

Using the parameters (average, standard deviation and regression coefficients) obtained for October 2001 – September 2004 (Table 2 and Table 3) in multiple regression on verification period (October 2004 – September 2005) R^2 is good.

Statistics on regression vs day-0 PM10			regression coefficients								
			day-0				day-1				day-2
R^2	explaned variance	RMSE	wind velocity	hmix	standford	total precip.	PM_{10}	wind velocity	hmix	total precip.	PM_{10}
0.72	0.89	0.65	-0.23	-0.19	0.08	-0.18	0.55				
0.65	0.77	0.74	-0.23	-0.28	0.07	-0.15		-0.10	-0.02	-0.20	0.30

Table 4: statistical scores and coefficients for the multilinear regression with log-normal variables for independent set.

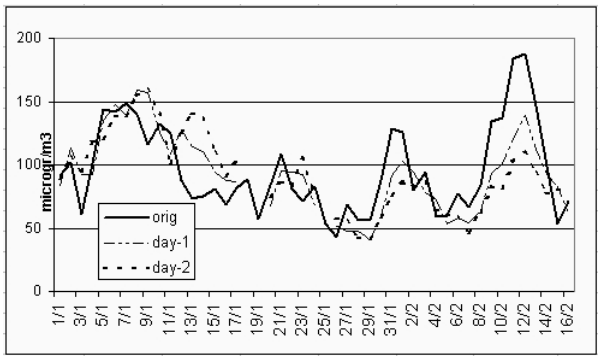

Figure 3: a month of verification period: PM10 original and built with the regressions are shown.

4. CONCLUSION

The fact that the results of the log-normal regression, presented in Table 3, are not worst than the direct regression, presented in Table 1, gives us some confidence on the approach used.
The results in the Table 3 gives some hints on the key meteorological phenomena controlling the short-time trend of PM_{10} concentrations on a flat plain such as Pianura Padana, which appeared to be much more important than the emissions and therefore can be successfully used for the short time forecast.
In particular we focused the importance on Hmix for the same day and on wind and precipitations also on the days before. The performance of the model on the verification set gives some good information on the reliability of this approach to the short-time PM_{10} forecasting.

References

Scire J. et al*., 2000. A user's guide for the Calmet Meteorological Model.*

Stull R. B., 1998. *Introduction to boundary layer meteorology*, Kluwer Academic Publishers.

MATLAB Manual, 2000. Using Matlab (version 6), Using Matlab Statistics (version 6).

Cacciamani C. et al*.,* 2001. Operational meteorological pre-processing at Emilia-Romagna ARPA Meteorological Service as a part of a decision support system for Air Quality Management. *International Journal of Environment and Pollution*, 16, 1-6.