SOURCE APPORTIONMENT OF $PM_{2.5}$ IN URBAN AREAS USING MULTIPLE LINEAR REGRESSION AS AN INVERSE MODELLING TECHNIQUE

Bruce Denby

Norwegian institute for air research (NILU), Kjeller, Norway

Abstract: In many countries emissions of particulate matter from urban sources, such as traffic and domestic wood burning, can lead to high episodic concentrations. Though it is important for air quality management and exposure studies to understand the individual source contributions to these concentrations, the complexity of the urban environment does not always allow a clear separation of sources when using conventional monitoring techniques that measure particulate mass only. Chemical analysis of the particulates, combined with receptor modelling, is one method for determining source contributions but these do not provide direct information on emissions. Inverse modelling methods, that make use of both dispersion models and measurements, can in principle be used to determine emissions strengths and distributions. However, the urban environment is generally so complex and the number of observations so limited that most inverse modelling methods cannot be effectively applied. In this paper a straight forward inverse modelling method, using multiple linear regression, is described and applied. The method determines the optimal fit of the calculated source contributions using dispersion modelling, providing scaling factors for the individual source contributions. The method is applied to the urban area of Oslo for $PM_{2.5}$ in the winter of 2004 and the results of the inverse modelling are compared to independent receptor modelling. The method shows that the modelled source contribution from suspended road dust is underestimated by a factor of 7 – 10. For domestic wood burning the method shows an overestimate of the modelled source contribution by a factor of 2 - 3. These results are confirmed using independent analysis by receptor modelling. The methodology cannot distinguish directly between model or emission error and so further assessment of the model itself, and its uncertainty, is required before concrete statements concerning emission strengths can be made.

Key words: *Dispersion modelling, receptor modelling, multiple linear regression, particulate matter, inverse modelling, urban air quality, emissions, source apportionment*

1. INTRODUCTION

In many countries emissions of particulate matter from urban sources, such as traffic and domestic wood burning, can lead to high episodic concentrations. Though it is important for air quality management and exposure studies to understand the individual source contributions to these concentrations, the complexity of the urban environment does not always allow a clear separation of sources when using conventional monitoring techniques that measure particulate mass concentration only. Though dispersion models may be used for this purpose (e.g. Peace et al., 2004; Laupsa and Slørdal, 2003) some sources, such as domestic wood burning or suspended road dust, have a high uncertainty in their emission strengths and may not be well represented when using dispersion models.

Chemical analysis of the particulates, combined with receptor modelling, is one method for determining source contributions but these do not provide direct information on emissions. Inverse modelling methods, that make use of both dispersion models and measurements, can in principle be used to determine emissions strengths and distributions. However, the urban environment is generally so complex and the number of observations so limited that most inverse modelling methods cannot be effectively applied.

In this paper a straight forward inverse modelling method, using multiple linear regression (MLR), is described and applied. The method determines the optimal fit of the calculated source contributions using dispersion modelling, providing scaling factors (regression slopes) for the individual source contributions. These scaling factors can be interpreted in terms of emission correction factors or as indicators of model bias for individual sources. The method is applied to measured and modelled $PM_{2.5}$ concentrations in the city of Oslo, Norway, in the winter of 2004, from January to May. The major source contributions to $PM_{2.5}$ in Oslo include long range transport, wood burning for domestic heating and traffic related emissions (both exhaust and induced suspension from the road surface).

It is difficult to validate source apportionment when using just models, however, for a limited number of days during the study period 38 filter samples were also collected and chemically analysed at one of the monitoring sites in Oslo (Laupsa et al., 2008). These chemical analyses were subsequently used as input for a receptor model, using Positive Matrix Factorisation, to determine the source contributions at that one site. The results of this receptor modelling are used for comparison with the MLR.

2. METHODOLOGY

Inverse modelling using multiple linear regression

The aim of the inverse modelling carried out in this study is to provide an indication of the average contribution of the source sectors to the total observed $PM_{2.5}$ concentration. We consider that these sources are additive in the following manner:

$$C(x,y,t) = \sum_{i=1}^{n} c_i(x,y,t) \qquad (1)$$

where C is the total concentration and c_i indicates the contributions from the n source sectors contributing to the total $PM_{2.5}$ concentration. The true source contributions at any particular site are unknown from the total $PM_{2.5}$ mass concentrations and so we use a simple linear model to describe these given by

$$C_{obs}(x,y,t) = \sum_{i=1}^{n} a_i c_{\mathrm{mod}\,i}(x,y,t) + \varepsilon_i(x,y,t) \qquad (2)$$

where ε is the error and the coefficients a are time independent. Writing it in this way infers that each individual source contribution can be scaled but the factor a_i to minimise the error ε. We choose to use multiple linear regression (MLR) to achieve this, though other variational methods may also be applied. With MLR the factors a_i are determined by minimising the mean square error. In the context here MLR is applied without a bias offset, i.e. the intercept is forced to pass through 0. This implies that there are no missing sources or background contributions in the model. A similar approach has been applied (Fushimi et al., 2005) to determine benzene emissions from an industrial complex.

Given a number of observations in time and space, the factors a_i can be determined. In this study we deduce these factors for two situations. The first using daily mean concentrations from the four available stations for the entire 103 day winter period, and the second using 12 hourly means corresponding to the available receptor modelling results at the single station (RV4) for the 38 days when filter samples have been analysed.

To assess the uncertainty in the MLR, bootstrapping methods are used to provide standard deviations of the regression slope parameters. This uncertainty analysis indicates the robustness of the regression to the limited dataset available. Boot strapping involves the random sampling of the dataset, with replacement, and the recalculation of the regression parameters a_i for each realisation. 10000 realisations are used to determine the standard deviations.

The methodology is generally applicable to any set of sources but the following conditions will apply.

1. The contribution of the different source sectors should not be highly correlated. MLR will not be able to distinguish between the sources in such a case.
2. There is an assumption of linearity in regard to the modelled and real source contributions when using MLR. If this is not the case then MLR will not work optimally.
3. The methodology is best applied when the source contributions are of the same order of magnitude. The methodology will not provide useful results for a particular source when it is significantly smaller than the others.

Measurements of $PM_{2.5}$ in Oslo

High levels of both PM_{10} and $PM_{2.5}$ may be observed in Nordic countries for short term averages during winter. This is due to a combination of adverse meteorological conditions and enhanced emission from wood burning (Yttri et al., 2005) and suspended dust and salt from road traffic, through the use of studded tyres and salting (e.g. Normann and Johansson, 2006). In 2004, the air quality monitoring network in Oslo consisted of 10 monitoring stations (Figure 1). At four of these stations, hourly $PM_{2.5}$ concentrations were measured using TEOM instruments and these are used in the current analysis. These include one urban background station, Aker Hospital, and three traffic stations, Kirkeveien, Løren and RV4. The mean concentration of $PM_{2.5}$ during the 2004 winter period at the four stations was 13.9 μgm^{-3}. Daily mean values of $PM_{2.5}$ from a regional background site, Birkenes, are used as boundary conditions in the dispersion model calculations.

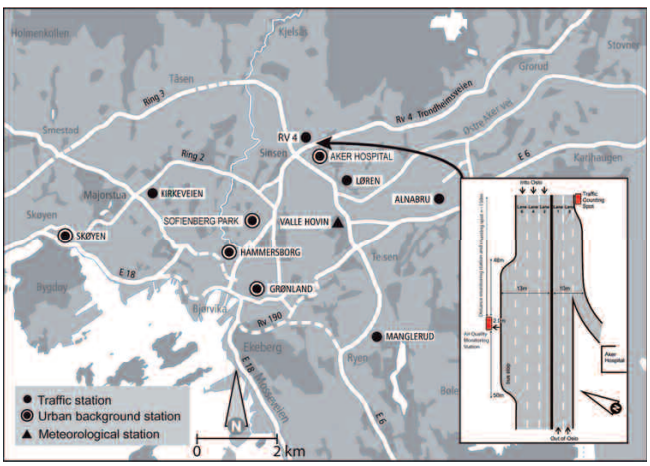

Figure 1. Location of monitoring stations in Oslo and a detailed description of the location of the RV4 station where the measurement campaign for the receptor modelling and source apportionment was carried out.

Dispersion modelling
The dispersion model employed is the AirQUIS-EPISODE model, which is an integral part of the AirQUIS air quality management system (AirQUIS, 2008; Slørdal et al., 2008). It is a Eulerian chemical transport model combined with sub-grid models for line sources, HIWAY-2 (Petersen, 1980), and point sources. For receptor points close to major roads the line source model is used to calculate the contribution from individual roads. Line sources more distant are included in the Eulerian grid model, as are other areally distributed sources such as domestic heating. AirQUIS-EPISODE uses a diagnostic wind field model, MATHEW (Sherman, 1978), based on data from a centrally located meteorological mast. For this study the model uses a 20 x 18 grid covering the Oslo region at a grid resolution of 1 km and using 10 vertical layers up to 2400 m with the lowest layer being 14 m. The model provides hourly concentrations at defined receptor points and these are converted to daily means for use in the regression analysis. For compatibility with the filter samples used for the receptor modelling 12 hourly mean concentrations are also calculated. For this study only primary emissions of $PM_{2.5}$ are considered with no chemical transformations or deposition occurring. The emission inventory includes all significant emission sources in Oslo, see Laupsa et al. (2008) for a more detailed description of the emissions.

Chemical analysis and receptor modelling
The chemical analysis and receptor modelling used for the comparison with the MLR was carried out for a limited period at the RV4 station, see figure 1, and this is described in detail in Laupsa et al. (2008). In summary, 78 12-hour filter samples of PM2.5, collected during two winter periods, were selected for chemical analysis. Based on the chemical analysis of the $PM_{2.5}$ samples receptor modelling was performed, two dimensional Positive Matrix Factorisation (Paatero, 1994), to detect and quantify the various source contributions. 38 of these days were available from the 2004 period and are used in this study.

3. RESULTS

Application to all stations for the 103 day winter period
For the regression carried out here only the sources associated with the regional background, traffic induced suspension, wood burning and other area sources are included. Industrial sources are insignificant in Oslo and are not included. $PM_{2.5}$ from traffic exhaust is not included as it is highly correlated, r^2=0.84 in the model, with the traffic induced suspension source. Traffic induced suspension was chosen, over traffic exhaust, for two reasons. Firstly $PM_{2.5}$ emissions from exhaust are better defined than those from suspension and secondly the receptor modelling carried out indicates a large discrepancy between the modelled and observed $PM_{2.5}$ contribution to traffic induced suspension. An alternative to choosing just the one traffic source is to lump them in a single source, however, this does not change the result of the regression to any significant extend.

MLR is carried out on all four stations simultaneously for the 103 day winter period. Any model source contributions not included in the MLR are subtracted from the total $PM_{2.5}$ concentration before the regression and added again when the regression model is calculated. The results of the regression in terms of the calculated regression coefficients, including their uncertainty, are shown in Table 1 for two different cases. The first where the regression factors are determined for all four selected sources, and the second where only the two most significant sources, i.e. wood burning and traffic induced suspension (road dust and salt), are fitted. In both cases the regression indicates that the model is overestimating the contribution from wood burning and significantly underestimating the contribution from traffic induced suspension. The source contributions averaged over the four stations and for the 103 day period are also shown in Figure 2.

Table 1. Multiple linear regression slopes determined for the various model sources of $PM_{2.5}$ for the 103 days in the period January 2004 - May 2004 as well as the 38 filter days (based upon daily mean values). All four available stations are included in the analysis. Uncertainty estimates show standard deviations of the slope parameters using bootstrapping methods.

Emission sources	103 day winter period		38 filter days	
	4 sources	2 sources	4 sources	2 sources
Regional background	1.22 ± 0.07		0.93 ± 0.06	
Traffic induced suspension	7.6 ± 1.0	8.6 ± 0.76	9.8 ± 1.7	7.8 ± 1.1
Wood burning	0.30 ± 0.06	0.29 ± 0.05	0.41 ± 0.12	0.33 ± 0.09
Other area sources	0.75 ± 0.42		-0.03 ± 0.83	

Scatter plots showing the results of both the model calculations and the MLR calculations are shown in figure 3 to indicate the effect of the application of MLR on the data. The results of the MLR for the 103 day period can be summarized as follows:

- The most dominant source in the regression is wood burning, having the highest coefficient of determination, followed by traffic induced suspension. There is little improvement when regional background contributions and other area sources are included in the regression, either in correlation or RMSE.
- Correlation (r^2) increases from 0.36 to 0.50 with the application of MLR and the RMSE decreases from 7.9 μgm^{-3} to 5.7 μgm^{-3}

- The uncertainty assessment in the regression slope indicates that the regional background, wood burning and traffic induced suspension are quite well statistically defined. There is a large uncertainty in the regression slope for the other area sources.
- The results for the entire measurement period (103 days) are also consistent with the 38 day filter period indicating that these 38 days are representative of the entire study period.

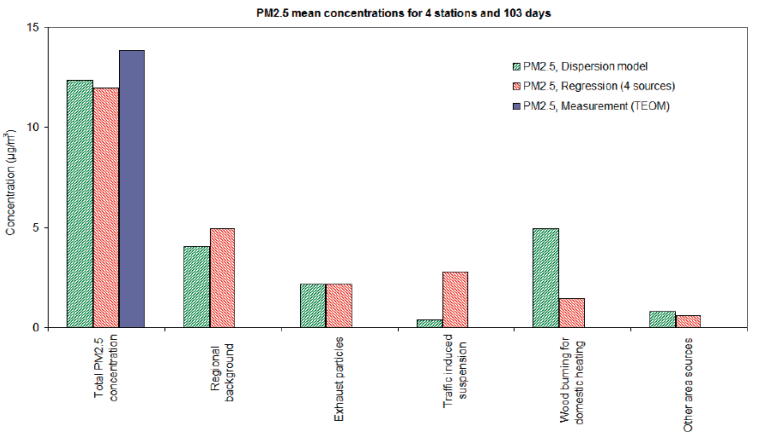

Figure 2. Estimated mean contributions to $PM_{2.5}$ (μgm^{-3}) from the different source categories using data from all four monitoring stations over the 103 day period. Shown is the dispersion model (green), the multiple linear regression (red) and the observed total $PM_{2.5}$ (blue).

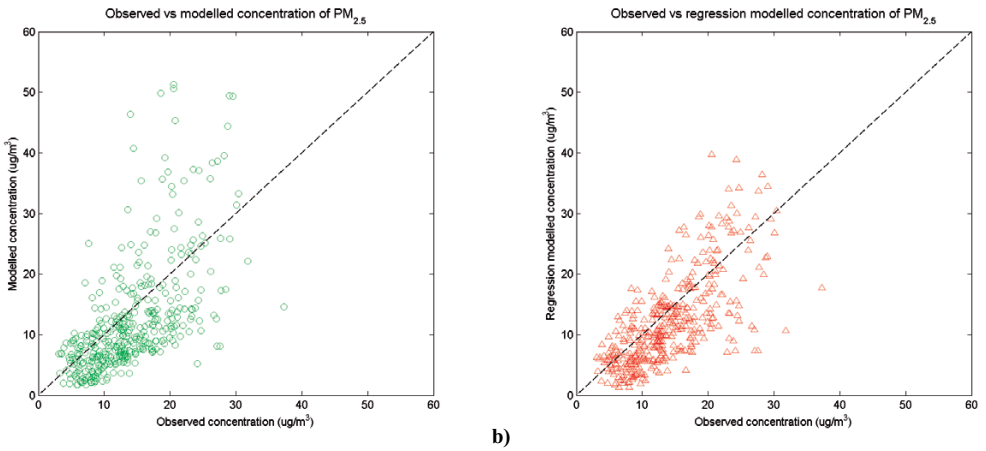

Figure 3. Dispersion model calculations verses observations of daily mean $PM_{2.5}$ using the four stations for the 103 day period. a) Model calculations without adjustment, b) Results after the multiple linear regression, as given in table 1.

Application to RV4 and comparison with receptor modelling results

In order to provide validation of the results, apart from the improved statistical parameters discussed previously, a limited number of days corresponding to the 38 filter days are compared to receptor modelling results. The comparison is shown in figure 4 below where the model, MLR and receptor model source contributions are shown. In this case only the two most correlated model sources, wood burning and traffic induced suspension, where included in the MLR due to the limited data available for the regression. In addition the receptor model could not distinguish between the traffic and other combustion sources so these have been lumped into the one source for the comparison. The receptor modelling confirms the differences found in the previous section for the 103 day period, i.e a significant under prediction of the traffic induced suspension and an over prediction of the wood burning contribution by the dispersion model.

The results of the MLR for the 38 filter days at the RV4 station can be summarized as follows:

- When only the two dominant sources are included then the regression slope for traffic induced suspension is found to be 10.6 ± 1.6. This is close to the factor of 7.1 found using the receptor modelling.

- When only the two dominant sources are included then the regression slope for wood burning is found to be 0.34 ± 0.22. This is slightly below the scaling factor of 0.54 found using the receptor model but within the determined uncertainty, which is quite large for the limited data set available.

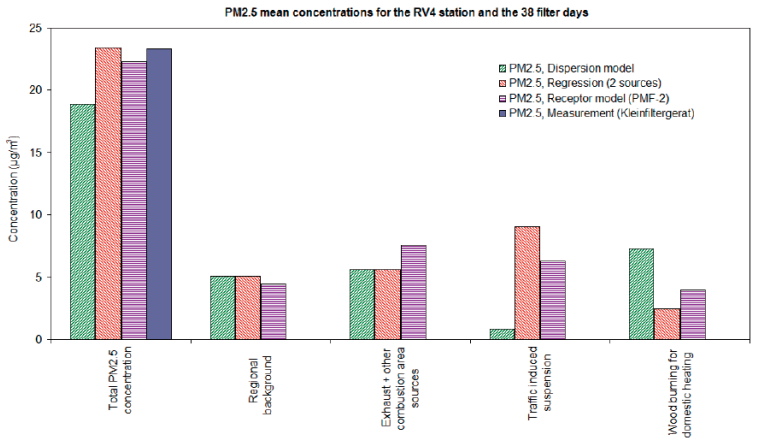

Figure 4. Estimated mean contributions to $PM_{2.5}$ (μgm^{-3}) from different source categories using only data from the RV4 station for the 38 filter days on which the receptor modelling is based. Shown are the dispersion model (green), the result after the multiple linear regression (red) and the corresponding results of the receptor modelling (purple). Measured total $PM_{2.5}$ concentrations are also shown (blue).

4. DISCUSSION AND CONCLUSIONS

Although the measured and modelled total $PM_{2.5}$ concentrations are, on the average, in good agreement at all sites in Oslo, MLR as an inverse modelling technique has shown large deviations for individual sources that compensate when combined together. The largest deviations are revealed for wood burning and traffic induced suspension where the optimal contributions differ from the dispersion model by a factor of 0.30 and 7.6, respectively. These have been qualitatively confirmed using chemical analysis and receptor modelling at one of the sites for a limited period.

The difference between the modelled and observed daily mean concentrations can be accounted for by either errors in the emission inventories, in the model formulation or in the meteorological input data. For the case regarding wood burning, which is modelled using the Eulerian grid model, it is not strictly possible to distinguish between these two uncertainties and it may well be that model formulation, e.g. vertical dispersion, initial emission heights or wind speeds are partly or wholly responsible for the differences found. Indeed, the days when measured concentrations are strongly over-predicted by the dispersion model due to wood burning contributions are also days characterised by measured wind speeds of < 2 ms^{-1}. Sensitivity studies concerning the vertical distribution of the wood burning emissions in the dispersion model also show large variations in model concentrations depending on the height at which emissions are introduced into the model grid. Another important source of uncertainty is the meteorological field generated by the diagnostic model, particularly in an urbanised area. Based on the current knowledge and available observational data, particularly meteorological, it is not possible to come to any firm conclusions regarding the cause of the differences found for the wood burning contribution.

For the case of traffic induced suspension the model, a Gaussian line source model, is less affected by uncertainty in meteorological conditions relating to dispersion or emission heights compared to the grid model. In this regard there is more confidence in the results of the line source model than the grid model. It should also be noted that the regression analysis cannot distinguish between exhaust emissions and traffic induced suspension due to the high correlation of these two sources in the model. In the results presented here we have assumed that the exhaust contribution is correct and any deviation is due to the traffic induced suspension. The receptor modelling results confirm this assumption, leading to confidence in the assertion that the emission estimate of traffic induced suspension is strongly underestimated.

It is important to have good knowledge of the various source contributions for the effective implementation of abatement strategies. This study has shown that the use of dispersion models, coupled with a simple inverse modelling technique (MLR) can be used to improve this knowledge. This improved knowledge can be used to update emission inventories or used as indicators of model weaknesses. Coupling the model to observationally based receptor modelling also provides important information and validation of the inverse modelling method.

REFERENCES

AirQUIS, 2008: The State of the Art Air Quality Management Tool: AirQUIS. URL: www.airquis.com.

Fushimi, A., Kawashima H., Kajihara H., 2005: Source apportionment based on an atmospheric dispersion model and multiple linear regression analysis. *Atmospheric Environment*, **39**, 1323-1334.

Laupsa, H., Slørdal L.H, 2003: Applying model calculations to estimate urban air quality with respect to the requirements of the EU directives on NO2, PM10 and C6H6. *Int. Journal of Env. Poll.,* **20**, 1-6.

Laupsa, H., Denby, B., Larssen, S. and Schaug, J., 2008: Source apportionment of particulate matter (PM2.5) using dispersion and receptor modelling. In press, *Atmospheric Environment*.

Norman, M., Johansson C., 2006: Studies of some measures to reduce road dust emissions from paved roads in Scandinavia. *Atmospheric Environment,* **40**, 6154-6164.

Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error-estimates of data values. *Envirometrics*, **5**, 111-126.

Peace, H., Owen B., Raper D.W., 2004: Comparison of road traffic emission factors and testing by comparison of modelled and measured ambient air quality data. *Science of the Total Env.*, 385–395.

Petersen, W.B., 1980: User's guide for Hiway-2: A highway air pollution model. U.S. Environmental Protection Agency, Research Triangle Park, NC. (EPA-600/8-80-018).

Sherman, C.A., 1978: A mass consistent model for wind fields over complex terrain. J. of Appl. Met., **17**, 312-319.

Slørdal, L.H., McInnes H., Krognes T., 2008: The Air Quality Information System AirQUIS. *Information Technologies in Environmental Engineering,* **1**, 40-47.

Yttri, K.E., Dye C., Slørdal L.H., Braathen O.A., 2005: Quantification of monosaccharide anhydrides by liquid chromatography combined with mass spectrometry: Application to aerosol samples from an urban and a suburban site influenced by small-scale wood burning. *J. of the Air & Waste Manag. Assoc.,* **55**, 1169-1177.