"Science Stays True Here"
Journal of Mathematics and Statistical Science (ISSN 2411-2518), Vol.5, 345-354 | Science Signpost Publishing

A Dynamic Factor Approach for Estimating an Aggregate Multipollutant Air Quality Indicator

Giuliana Passamani
Department of Economics and Management, University of Trento, Trento, Italy.
E-mail address: giuliana.passamani@unitn.it
Paola Masotti
Department of Economics and Management, University of Trento, Trento, Italy.
E-mail address: paola.masotti@unitn.it

Abstract

Measurement of air pollution has become an important issue since it has been established that air quality is closely connected to human health and environment. International organizations as well as local authorities are particularly concerned with air pollution, but, in spite of the huge amount of data on various pollutants recorded frequently at the monitoring sites located in many countries all over the world, the problem of properly synthetizing the available information is still a matter of discussion in the specialized literature. In this paper we present an explicit dynamic time series factor model that implicitly determines a variable which can be thought of as measuring the state of local air pollution. With the suggested operative approach, we aim to contribute to measuring air quality, by proposing a methodological procedure leading to the estimation of a single site indicator determined jointly by present and past pollution as well as by the meteorological conditions. These single indicators are then spatially aggregated using principal component analysis. The advantage of using this dynamic factor model for the empirical analysis is that, besides measuring air pollution, we can use the estimated model for forecasting future air pollution, given the meteorological predictions. The application of the model in the present paper considers a pollution data set collected at different monitoring sites in the alpine province of Trento[1].

Keywords: Air pollution measurement, Aggregate air quality indicator, Dynamic factor model.

1. Introduction

Since the end of past century, reducing air pollution and improving the quality of environment have been among the main efforts of international organizations. These have been pursued through increasing information on the consequences of pollution and by working towards the introduction of environmental laws and directives, in order to establish new and appropriate regulations. For this purpose, European Union legislation defines evaluation and management methods for air quality and set the standards for the monitoring networks. Moreover, the European Environment Agency's European Air Quality Index (AQI), calculated by local agencies for single monitoring sites, allows interested people to understand more about air quality where they live. The index is based on concentration values for up to five key pollutants, including particulate matter (PM_{10}), fine particulate matter ($PM_{2.5}$), ozone (O_3), nitrogen dioxide (NO_2) and sulphur dioxide (SO_2). It considers the potential impact of air pollution on health, impact given by the pollutant for which concentrations are poorest in terms of health effects.

[1] The data set for the empirical analysis has been provided by “Agenzia Provinciale per la Protezione dell’Ambiente (APPA)” of the Province of Trento (Italy).

The main concern about European AQI lies in the fact that it doesn't take into account the simultaneous conjoint effect on health of multiple air pollutants. Given this concern, some specialized literature has devoted particular effort towards calculating an aggregate air pollution index based on the coexistence of concentrations of various pollutants that, even at lower values, can affect significantly the health of people exposed to them. A certain number of air quality indices have been therefore proposed, combining observations on a variety of pollutants at different monitoring sites over a defined area, in order to calculate a simple indicator summarizing air quality, as in Bruno and Cocchi (2002), where they obtain a synthetic index value by means of hierarchical median-maximum aggregation processes, or as in Plaia, Di Salvo, Ruggeri and Agrò (2013), where they suggest an index based first on a spatial aggregation and then on a pollutant synthesis.

The aim of the present paper is to propose a statistical model, for analysing multiple pollutant time series, that has already been applied successfully for computing leading economic indicators summarizing the state of macroeconomic activity (Stock, 1989). This statistical model, the dynamic factor model, provides a framework within which we can calculate a synthetic measure of air pollution at a given site by means of a stochastic model where we combine, in a dynamic contest, the daily measurements of air pollutants with the meteorological conditions and the past observed pollution level. That is, we suggest a stochastic dynamic procedure for synthetizing in a single indicator the state of local air pollution. Given that meteorological conditions can change within a distance of some miles, the daily concentration values for each pollutant cannot simply be aggregated over different monitoring sites, but, first of all, we must analyse and aggregate the various pollutants at the same monitoring site for which we have also the observations on meteorological variables. Once the synthetic air pollution state measure has been calculated for a given site, an aggregate air quality indicator can be calculated for an entire area of interest. In the present paper the aggregate indicator is obtained by means of principal component analysis.

The plan of the paper is the following. In Section 2 we present the characteristics of the available pollution data set. In section 3 we discuss the methodological approach and the dynamic factor model adopted for analysing the available daily time series data set. In Section 4 we describe the results in terms of the estimated pollution state indicator for each site and then we aggregate them over the different monitoring sites covering the area of interest. In Section 5 we conclude the paper and outline possible lines for further research.

2. The Pollution Data Set

The set of raw data available for the empirical analysis consists of hourly observations on few pollutants recorded at different monitoring sites covering the area of the province of Trento, in Northern Italy. The hourly observations extend for a time period of two years, 2014 and 2015.

In order to have homogeneous data on the same pollutants for a certain number of monitoring sites, we have to focus just on three of them: particulate matter, PM_{10}, nitrogen dioxide, NO_2 and ground-level ozone, O_3. This is a constrained but reasonable choice since it's well acknowledged by international organizations that exceedances of air pollutants pose serious health risks and PM_{10}, NO_2 and O_3 are generally recognized as the three chemicals that most significantly affect human health. These chemicals are observed at different monitoring sites located in traffic and non-traffic areas. In addition, for the same sites we have the availability of hourly observations on some meteorological variables of which we retain that wind speed and

precipitations can have a significant impact on air pollution.

The first step of our analysis consists in transforming hourly time series observations, obtained from continuous measurements of each of the three pollutants, in daily observations, as follows: the daily values on PM_{10} correspond to the 24-hour running means, while the values on NO_2 and O_3 are the maximum hourly concentrations within the single day. Their unit of measurement is $\mu g/m^3$. For the meteorological conditions, wind speed (m/s) and precipitations (mm) are taken as daily means.

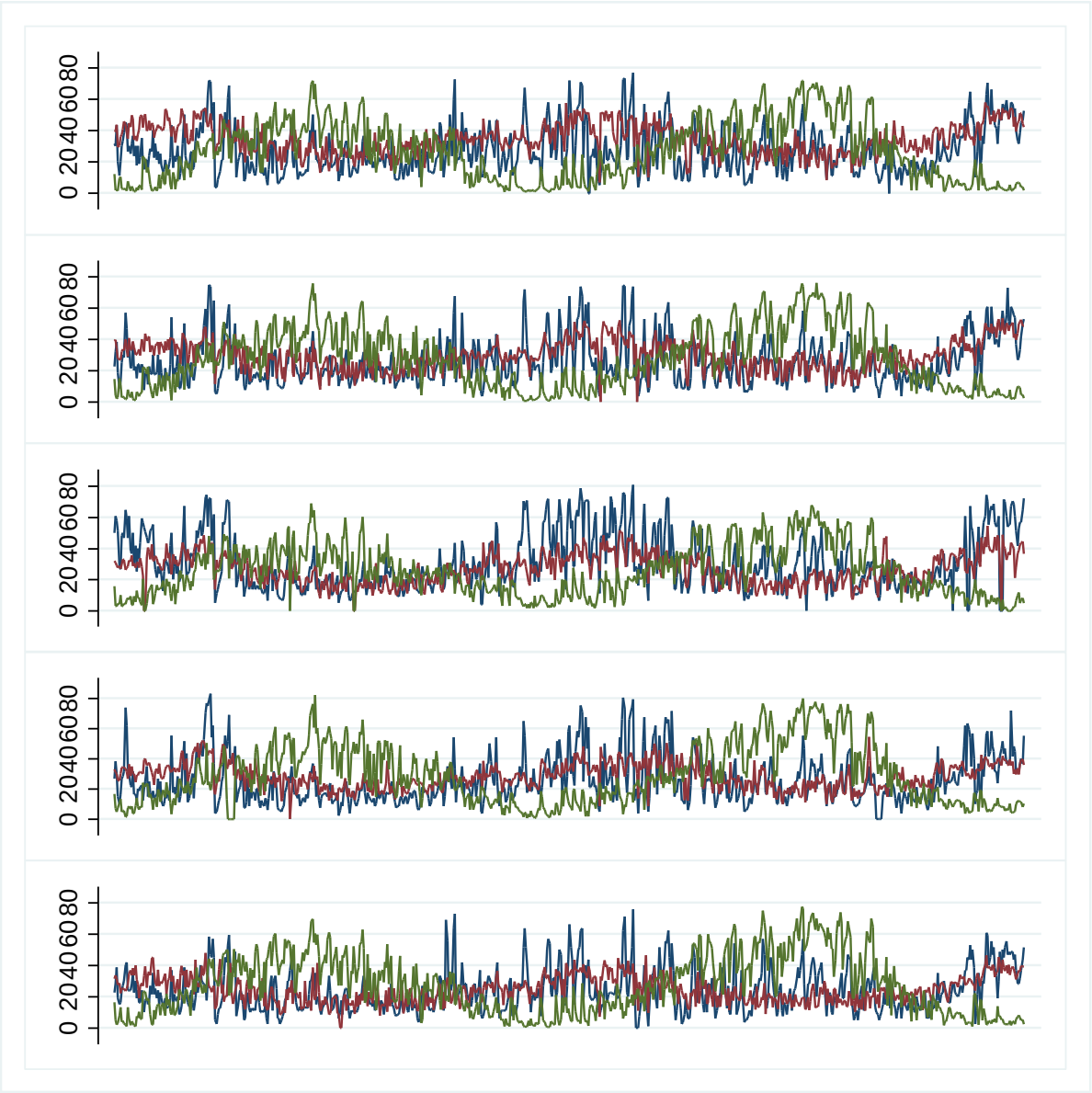
In order to be able to study and compare the evolution over time of the selected pollutants, characterized by different orders of magnitude, we need to standardize[2] them by computing new daily values calculated by the segmented linear principle as in Murena (2004), where the resulting standardized values will be in the range [0, 100] for each pollutant, with the following comparable upper breakpoint values: 25 for "good air quality"; 50 for "low pollution"; 70 for "moderate pollution"; 85 for "unhealthy for sensitive groups"; 100 for "unhealthy". Hadn't we standardized the observations, we could not have the same breakpoint values for any of the three pollutants.

In Figure 1 are represented the standardized daily values for the three pollutants for five selected monitoring sites. The sites refer to two main urban areas, Trento and Rovereto, to an area with quite a significant road traffic, Borgo Valsugana, to a touristic area not directly influenced by urban sources, Riva del Garda, and to a suburban area suffering for some road traffic, Piana Rotaliana.

As can be noticed, the pollutants are characterized by a clear seasonal variation, with PM_{10} and NO_2 positively correlated and both negatively correlated with O_3. Moreover, most days, for each pollutant, can be classified in the categories "low pollution" and "moderate pollution", but there are also many days "unhealthy for sensitive groups", for PM_{10} - mainly, Borgo Valsugana - and for O_3.

[2] The standardization procedure is particularly recommended by US Environmental Protection Agency (2006).

Figure 1. Standardized daily observations, over 2014 and 2015, for PM10, represented in blu, NO2, in red e O3, in green, for each of the selected monitoring sites: in order, Trento, Rovereto, Borgo Valsugana, Riva del Garda and Piana Rotaliana



3. The Methodological Approach

After the brief descriptive presentation of the pollution data, we focus on the proposed statistical methodology aiming at the construction of an indicator measuring the state of pollution. The suggested stochastic dynamic model for analysing the observed multiple time series is based on the principle that few unobservable dynamic factors drive the co-movements observed in a higher dimensional vector of endogenous time series variables which can also be affected by exogenous covariates, as well as by a vector of mean-zero idiosyncratic disturbances. For the purpose of our methodological approach, we assume the existence of a single unobservable dynamic factor underlying the observed pollutants, representing the state

of pollution.

The specification of the stochastic dynamic factor model that represents the framework for the empirical analysis consists of a first stochastic equation[3]:

$$\mathbf{y}_t = \boldsymbol{\lambda} f_t + \boldsymbol{\beta}' \mathbf{x}_t + \mathbf{u}_t, \quad (1)$$

where $\mathbf{y}_t$ denotes the $(k \times 1)$ vector of observed endogenous variables, k being the number of pollutants analysed, and $\mathbf{x}_t$ the $(m \times 1)$ vector of observed exogenous variables, m being the number of meteorological variables taken into account; f_t denotes the unobserved common factor that represents a synthetic measure of air pollution state, and $\mathbf{u}_t$ is a $(k \times 1)$ idiosyncratic disturbance vector. The dynamic properties of the unobserved common factor are implied by the following autoregressive process of order one:

$$f_t = \psi f_{t-1} + \nu_t, \quad (2)$$

where ν_t is a scalar disturbance. The idiosyncratic disturbances of equation (1) are assumed to be uncorrelated with the factor disturbance at all leads and lags.

Moreover, the idiosyncratic disturbance terms are assumed to be generated by a vector autoregressive process of order one as follows:

$$\mathbf{u}_t = \boldsymbol{\Phi} \mathbf{u}_{t-1} + \boldsymbol{\varepsilon}_t, \quad (3)$$

where $\boldsymbol{\varepsilon}_t$ is a $(k \times 1)$ vector of disturbances.

The unknown parameters are $\boldsymbol{\lambda}$, a $(k \times 1)$ vector containing the unknown dynamic factor loadings; ψ, the autoregressive factor parameter; $\boldsymbol{\Phi}$, a $(k \times k)$ matrix of autoregressive parameters for the disturbances. Model (1), (2) and (3) can thus be considered as a stochastic dynamic model with vector autoregressive errors, where the conditional mean of the unobserved factor is assumed to vary over time according to an AR(1) model.

The first issue at hand is to estimate the factor. With the further assumption that the disturbances making up the vector $\boldsymbol{\varepsilon}_t$ are i.i.d. N(0, σ_i^2), $i = 1, ..., k$, the estimation is performed using a maximum likelihood approach, implemented by writing the model in a linear state-space form, where the state equations represent the evolution of the unobserved common factor and of the idiosyncratic disturbances, while the space or measurement equation relates the observed variables to the unobservable ones. The Kalman filter is used to compute the Gaussian likelihood function in prediction-error form: the filter recursively obtain conditional means and variances of both the unobserved states and the measured dependent variables that are used to compute the likelihood. This approach unites the statistical efficiency of the state-space approach with the robustness and convenience of the principal component factor analysis approach: state-space/Kalman-filter estimates can produce substantial improvements in the estimates of the factors and common components if the time behaviour of the common component is persistent, as in the present case. It is designed to handle also cases in which the number of modeled endogenous variables, k, is

[3] A similar model was used by Fontanella et al. (2007) for the analysis of environmental pollution in the Milan district, following the work of Forni et al. (2000).

small and a further advantage of this parametric state-space formulation is that it can manage data irregularities[4]. Once the model has been estimated, it could be used in order to predict the unobserved common variable $\hat{f}_t$ for each monitoring site. The prediction method estimates the states at each time by a Kalman smoother and using all the sample information.

The predicted unobserved common variables $\hat{f}_t$ for each site can then be spatially aggregated in a single area pollution indicator by using conventional principal component analysis performed on the covariance matrix of the unobserved common variables. The scores obtained considering only the first principal component represent the spatial synthesis of air pollution, that is the overall multipollutant air quality indicator for the geographical area of interest.

4. The Estimated Site Air Pollution Indicators and the Overall Air Quality Indicator

Given the methodological approach adopted for the empirical analysis, data need not to be pre-processed in order to standardize pollutants with different orders of magnitude, as has been done in Section 2 for comparing the various pollutants. The standardization procedure must be applied when the purpose of the analysis is to calculate air quality indices by means of some aggregation functions, as in a large part of the literature has been done (see, inter alia, Murena, 2004, Plaia et al., 2013, Li et al., 2014).

For each of the selected monitoring site we have estimated the model made up by equations (1), (2) and (3). The estimates show that the unobserved common factor $\hat{f}_t$ is quite persistent and it's a significant predictor of the observed variables, that is the factor loadings are highly significant and they show a positive sign[5], indicating that the three pollutants contribute together to determining the pollution indicator.

The two meteorological variables that we assume to determine the level of pollutants, show different effects: wind affects negatively and significantly PM_{10} and NO_2, but positively and significantly O_3 in any monitoring station; precipitations affect significantly and negatively O_3 in Trento and Rovereto, significantly and negatively O_3 and significantly and positively NO_2 in Piana Rotaliana, while they have no significant effect in Borgo Valsugana, and significant and negative effects on PM_{10} and NO_2 and significant and positive effect on O_3 in Riva del Garda. Briefly, wind has a clear impact everywhere, while precipitations have not.

In order to have a better understanding of the results obtained with the suggested methodological procedure, we make use of a combined graph for the selected monitoring sites. In Figure 2 we represent with a thicker orange line the estimated common dynamic factor for each site. In the same combined graph we can observe, for each site, the standardized measurements of the three pollutants whose linear combination, together with the weather conditions, determines the unobserved common factor that is the site indicator measuring the state of air pollution. Given the differences in the order of magnitude of the three pollutants, we have to use standardized data obtained using the linear interpolation method if we want to compare the estimated pollution state indicator with the observed air pollutants. It's rather surprising to note how the procedure has been able to summarize pretty well, in all cases, the daily observations with a single smooth indicator.

[4] For a detailed presentation of the methodology, we advise to refer to Stock and Watson (2011).

[5] Results are available upon request. They have been obtained using software Stata 13

In Figure 3 we represent the estimated air pollution state indicators for the five monitoring stations in one graph. It's interesting to note the seasonal variations shown by the indicators over the two years of observation. As expected, the two urban monitoring sites, Trento and Rovereto, show higher air pollution levels with respect to the other sites. Borgo Valsugana seems to suffer from pollution less than the two urban areas, though, according to official data recorder by APPA, it is affected by pollution even in a worse manner than the others. APPA is the provincial environment protection agency that collect data on pollutants and calculate the air quality indices (AQIs) in accordance with the European and national directives. These calculated AQIs are then transmitted to the European

Environmental Agency that make them publish on its website[6]. If we look at those indices, we can see that Borgo Valsugana is worse affected mainly by particulate matter of a smaller dimension, $PM_{2.5}$, a chemical primarily due to road transport. Data on $PM_{2.5}$ are not available for all the sites and for this reason this has not been taken into consideration in the empirical analysis.

The overall pollution state indicator for the province of Trento has been obtained by performing principal component analysis as described in Section 3. The time series represented in Figure 4 corresponds to the scores on the first principal component (PC) that explains more than 94.0% of the total variance of the five analysed site indicators, a very high percentage. Moreover, the loadings/weights of each site on the first PC appear to vary within quite a narrow range, that is between 0.37 and 0.53. In the graph the red horizontal lines represent the breakpoint levels defining the pollution categories mentioned in Section 2. As can be noticed, most days can be classified in the categories "low pollution" and "moderate pollution", but there are many days "unhealthy for sensitive groups" in winter time.

[6] EEA: https://www.eea.europa.eu/themes/air/air-quality-index

Figure 2. The standardized pollutants and the estimated pollution indicator $\hat{f}_t$, in orange, at each selected monitoring site

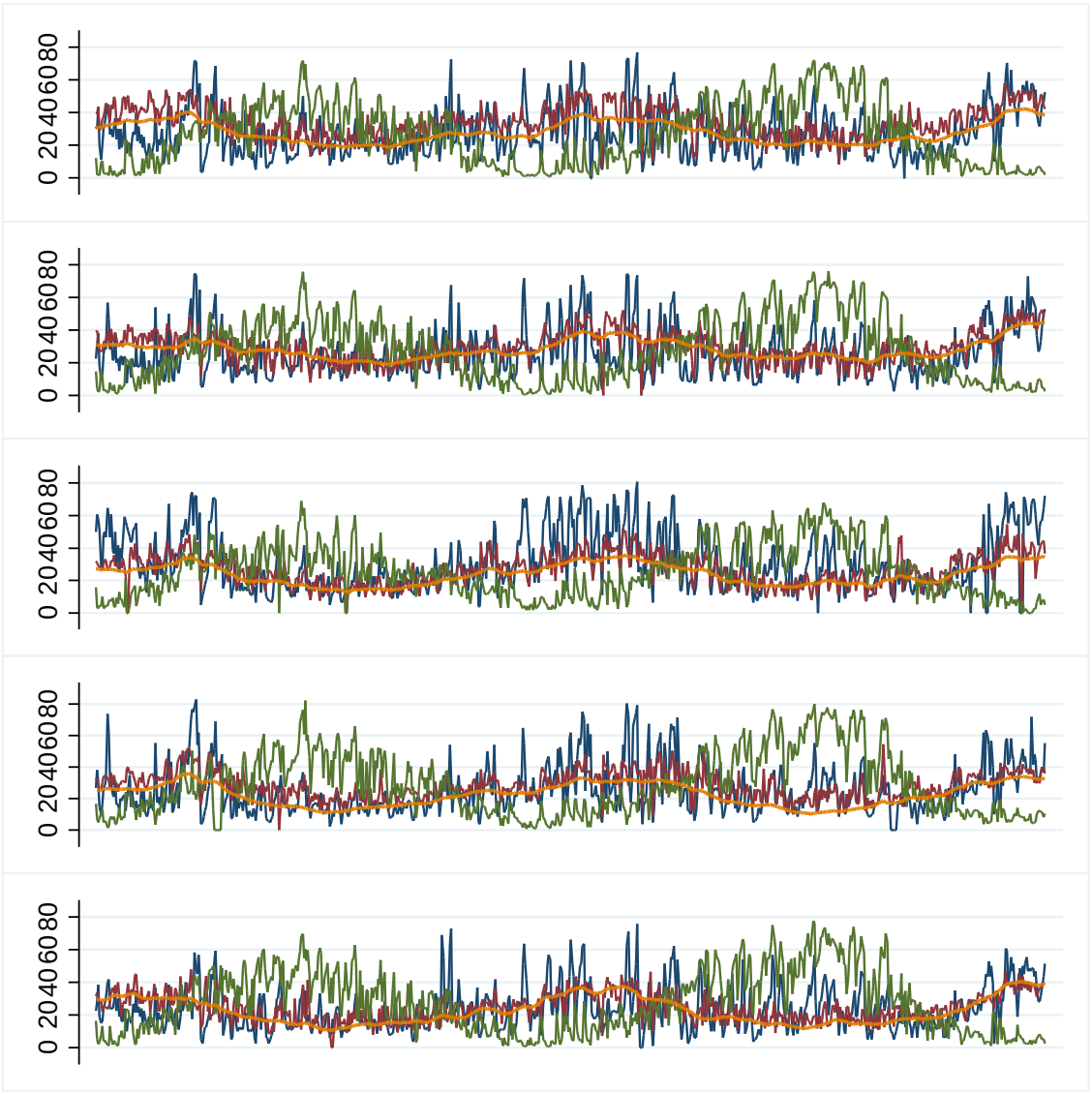


Figure 3. The estimated pollution state indicators for the different monitoring sites

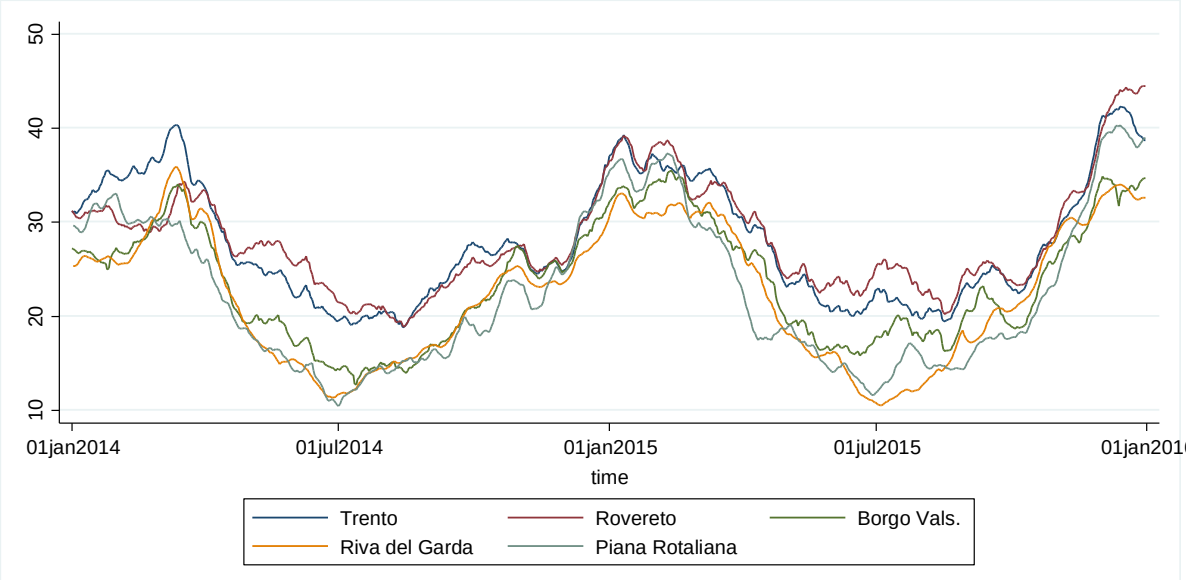
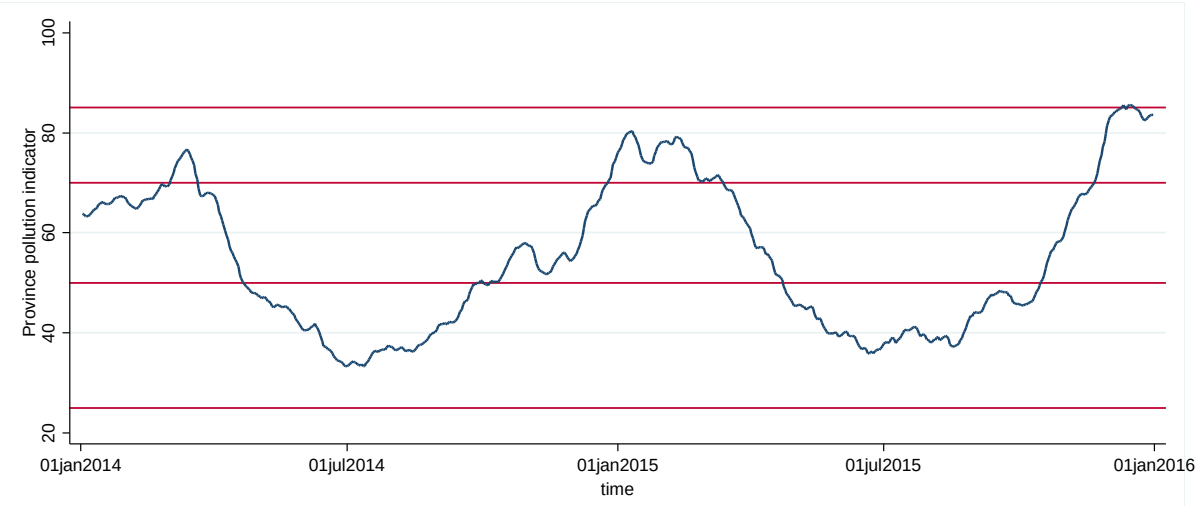


Figure 4. The estimated pollution state indicator for the province of Trento



5. Conclusions

As already mentioned, for reasons due to the availability of data, our empirical analysis takes into account just three main pollutants: for sure, more interesting results could be obtained if we had data even on sulphur dioxide, SO_2, and carbon oxide, CO, as well as on $PM_{2.5}$, in such a way to be able to compute a synthetic indicator that better depicts the situation of air quality.

Anyway, it's to be underlined that the purpose of this paper is principally the proposal of a statistical procedure to be applied for analysing multisite multipollutant data within a dynamic framework, and not just to simply calculate air quality indices, which is already the aim of much work in the specialized literature. The advantage of the dynamic-factor model used for the empirical analysis has been highlighted, but further research could be done, particularly in the direction of being able to better forecasting future air pollution and its variability, given also the predicted weather conditions.

Another appealing further issue would be the suggestion of a proper procedure for aggregating the

estimated air pollution indicators in a single one for an entire geographical area, aggregation that in the present paper has been obtained using PC analysis. Even if the aggregation procedure chosen has led to the estimation of a realistic aggregate indicator, we think that more research work should be done in this direction. Moreover, this issue could be of particular interest especially in the case we want to synthetize in a single reliable measure the pollution data collected by means of air monitoring networks covering a large area with similar characteristics, like a metropolitan area.

References

[1]. Bruno, F. and Cocchi, D. (2002). A unified strategy for building simple air quality indices. Environmetrics, 13, 243–261.

[2]. Fontanella, L., Ippoliti, L. and Valentini, P. (2007). Environmental Pollution Analysis by Dynamic Structural Equation Models. Environmetrics, 18, 265-83.

[3]. Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000). The Generalized Dynamic-Factor Model: Identification and Estimation. The Review of Economics and Statistics, 82, 540-54.

[4]. Li, L., Qian, J., Ou, CQ., Zhou, YX., Guo, C. and Guo, Y. (2014). Spatial and temporal analysis of Air Pollution Index and its timescale-dependent relationship with meteorological factors in Guangzhou, China, 2001–2011, Environmental Pollution, 10, 75-81.

[5]. Murena, F. (2004). Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naples. Atmospheric Environment, 38, 6195-6202.

[6]. Plaia, A., Di Salvo, F., Ruggieri, M. and Agrò, G. (2013). A Multisite-Multipollutant Air Quality Index. Atmospheric Environment, 70, 387-391.

[7]. Stock, J.H. and Watson, M.W. (1989). New Indexes of Coincident and Leading Economic Indicators. NBER Macroeconomics Annual 1989, 351-393.

[8]. Stock, J.H. and Watson, M.W. (2011). Dynamic factor models. Chapter 2 in Clements, M. J. and Hendry, D. F. (Eds.), Oxford Handbook of Economic Forecasting, OUP, 35-59

[9]. US Environmental Protection Agency (2006). Guidelines for the Reporting of Daily Air Quality – the Air Quality Index (AQI). U.S. EPA Office of Air Quality Planning and Standards Research, Triangle Park, North Carolina.