International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:9, No:2, 2015

Air Pollution and Respiratory-Related Restricted Activity Days in Tunisia

Mokhtar Kouki Inès Rekik

***Abstract*—**This paper focuses on the assessment of the air pollution and morbidity relationship in Tunisia. Air pollution is measured by ozone air concentration and the morbidity is measured by the number of respiratory-related restricted activity days during the 2-week period prior to the interview. Socioeconomic data are also collected in order to adjust for any confounding covariates. Our sample is composed by 407 Tunisian respondents; 44.7% are women, the average age is 35.2, near 69% are living in a house built after 1980, and 27.8% have reported at least one day of respiratory-related restricted activity. The model consists on the regression of the number of respiratory-related restricted activity days on the air quality measure and the socioeconomic covariates. In order to correct for zero-inflation and heterogeneity, we estimate several models (Poisson, negative binomial, zero inflated Poisson, Poisson hurdle, negative binomial hurdle and finite mixture Poisson models). Bootstrapping and post-stratification techniques are used in order to correct for any sample bias. According to the Akaike information criteria, the hurdle negative binomial model has the greatest goodness of fit. The main result indicates that, after adjusting for socioeconomic data, the ozone concentration increases the probability of positive number of restricted activity days.

***Keywords*—**Bootstrapping, hurdle negbin model, overdispersion, ozone concentration, respiratory-related restricted activity days.

I. Introduction

IN the past 40 years, research on air quality has increasingly become an issue of great concern in view of the accumulated evidence that demonstrates adverse effects of air pollution on human health. The assessment of pollution external costs is of great interest to policy makers, since it allow them to integrate the environmental issues in decision making. A variety of methods has been used to the valuation of these costs. A common approach used to evaluate the costs of air pollution is the Concentration-Response Relationships (CRRs).

This method is based on the estimation of the physical impact of socio-economic and environmental variables, such as ambient concentrations of air pollution, on the morbidity. The most used indicators of morbidity are days of work loss (WLD), days of restricted activity (RAD), and days of respiratory-related restricted activity (RRAD).

M. Kouki is with Université de Carthage, Ecole Supérieure de la Statistique et de l'Analyse de l'Information, Ecole polytechnique de Tunisie, LEGI, B.P. 748, 2078 La Marsa, Tunis, Tunisie (e-mail: mokhtar.kouki@essai.rnu.tn).

I. Rekik is with Université de Tunis El Manar, Institut Supérieur d'Informatique, Faculté des Sciences Economiques et de Gestion de Tunis, LAREQUAD, B.P. 248 El Manar II 2092 Tunis, Tunisie (phone +216-97-477483 ; fax : +216-71-706698 ; e-mail: i.rekik@yahoo.fr).

Several studies have identified CRRs for US and some developed countries data, using the number of RRADs as a measure of health impact of air pollution and have shown a significant relationship between air pollution and human health [1]-[4]).

Over the last decades Tunisia has recognized a fast development of the principal causes of air pollution. The annual growth rate of the urban population was 1.824% between 1995 and 2005. Increasing rates of industrialization and rising demands for motor vehicles have been registered. The annual growth rate of the motor vehicles was more than 6% over the same period. These factors are associated to an intensive use of combustible fuels accompanied by a poor environmental regulation. The situation is implying an increasing pressure on the public health.

In this paper, we investigate the morbidity effects of air pollution in Tunisia by estimating CRRs based on Tunisian health survey data. The data are collected from 407 randomly chosen individuals from 9 Tunisian zones. These zones are geographically close to the air pollution monitors planted by the National Agency for Environmental Protection in Tunisia. Socioeconomic data, air quality measures, temperature and respiratory-related restricted activity days are collected. Air quality measures are provided by air pollution monitors and are attributed to delegations that are less than 10 kilometers away. Our sample is drawn from individuals who live in these areas. Individuals' exposure to air pollution is measured using data from the air pollution monitors nearest their residences. As we are concerned by the health status of adults, respondents are aged between 18 and 60 years. Following [1], we exclude older peoples from our sample because for these persons respiratory problems are always likely to be notably related to their age. Let us notice that the number of exposed individuals are about 1 600 000 individuals. The survey period is from March to April 2006.

The rest of the paper is organized as follows. In section II, we expose our statistical methodology. Section III deals with data and exploratory statistics. Econometric results are presented in Section IV. Section V is dedicated to results discussion and Section VI concludes.

II. Methodology

Overdispersion in count data models can be considered as a direct consequence of unobserved heterogeneity. In this context, [5] developed continuous mixture models for unobserved heterogeneity. These models consider that the conditional distribution of the dependent variable given an unobserved heterogeneity variable ν is Poisson with

parameter $\lambda\nu$. The negative binomial [1] (Negbin) is particular case with ν a Gamma variable of mean $E(\nu)=1$ and variance $V(\nu)=\alpha$.

For the zero-excess outcome, [6] suggested modified count models in which zeros and nonzero are not governed by the same data-generating process. These models are the zero-inflated and the hurdle model which have been widely used in the literature [7], [8], [5]. We consider that these models are well adapted in our context. Indeed, as we will show below, the sample data contains two sub-samples. The first sample contains the individuals who are in good health and never had respiratory problems and the second sample contains those who are not healthy. Even the latter induces a high percentage of zero RRADs. To handle with this phenomenon, the zero inflated model (ZIP) specifies that the probability of the zero outcomes is additively augmented by a logistic function φ_i (see Table I which summarizes count data models used in the empirical illustrations.).

The hurdle model allows the possibility that the mechanisms that determine the likelihood of respiratory problems may differ from the mechanisms that determine their duration. Then, for our study, we need to specify the probability π_0, of being healthy in the sample period (RRAD=0) and the distribution for the positive number of RRADs.

Another alternative to deal with overdispersion is the finite mixture (FM) models. The FM formulation uses a discrete representation of unobserved heterogeneity. It supposes that the density $f(y_i)$ of Y_i is:

$$f\left(y_i\right)=\sum_{j=1}^{g}\pi_j f_j\left(y_i\right), \qquad (1)$$

$$0<\pi_j<1,\quad j=1,...,g \text{ and } \sum_{j=1}^{g}\pi_j=1.$$

Equation (1) defines a g-component finite mixture density. Each term in the sum is the product of mixing probability (weight) π_j and the component density $f_j\left(y_i\right)$. To fit our data we use a Poisson mixture model which considers that the component distributions are Poisson distributions (line 6 of Table I). The mixing proportions π_j are commonly parameterized using the logistic function:

$$\pi_j\left(x_i,\beta\right)=\frac{\exp\left(x_i'\beta_j\right)}{1+\sum_{h=1}^{g-1}\exp\left(x_i'\beta_h\right)},\quad j=1,...,g-1. \qquad (2)$$

$$\pi_j\left(x_i,\beta\right)=1-\sum_{h=1}^{g-1}\pi_h\left(x_i,\beta\right).$$

[1] Reference [9] considered the Negbin models and demonstrated the superiority of the Negbin2, for which the conditional variance function is quadratic in the mean. This Negbin implementation is adopted in what follows.

The parameters of the mixed models are estimated using an EM algorithm for mixture model estimation [10]. In practice, g should also be estimated. In general, the g components of the mixture correspond to g different groups identified throw some observable characteristic(s). For g varying between 1 and the sample size n, the mixture approach is considered in [11] as a semiparametric compromise between the fully parametric model for $g=1$, and a nonparametric model given in the case of $g=n$ by the Kernel method of density estimation[2].

The question of the number of components in a mixture model is always approached in terms of an assessment of the smallest number of components in the mixture compatible with the data [11]. Reference [13] suggested a selection approach consisting of two stages. At the first stage, we fix an upper bound for g, this value that we set g^* is often a small number like 2, 3, or 4. Saturated mixture models (containing all possible covariates) are estimated for all values of $g\le g^*$. Then we select g that minimizes information criteria. The widely used criteria are the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC).[3]At a second stage, likelihood ratio tests are used to compare between nested g-component mixture models, when the goal is inference about model parameters.

Compared with samples used in similar studies, our sample is relatively limited in size. This is due to practical difficulties to interview a large number of individuals. In order to improve the quality of the data and make them more representative of the mother population, post-stratification[4] and bootstrapping techniques are used. Post-stratification classifies the sample into post-strata according to different characteristics and then provides a set of sub-samples (empirical post-strata) from which pseudo-samples are drawn respecting population frequencies. The samples obtained are then used for parameters estimation and statistical testing. This approach is based on the bootstrapping techniques.

III. Data and Exploratory Statistics

As indicated below, the data are elicited using a questionnaire surveying individuals from different regions in Tunisia for which air pollution is monitored. The number of RRADs experienced during the 2-week period prior to the date of each interview was to be reported by each respondent.

[2] The nonparametric approaches identify the number of components by the resulting placement of modes in the density estimate. This approach was used in [12] when testing for multimodality using nonparametric kernel density estimation technique.

[3] The Monte Carlo study in [13] shows that the BIC is more reliable than the AIC. Indeed, the BIC choose always the correct models. The AIC choose the correct model in 96% of the time.

[4] See [14] for a detailed description of post-stratification.

TABLE I
POISSON AND OVERDISPERSED COUNT DATA MODELS

Model	$f(y_i/x_i)$	$E(y_i/x_i)$	$V(y_i/x_i)$
Poisson	$\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$, $y_i = 0, 1, 2, \ldots$	$\mu_i = \exp(x_i'\beta)$	μ_i
Negbin2	$\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i+1)}\left(\frac{\alpha^{-1}}{\mu_i+\alpha^{-1}}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\mu_i+\alpha^{-1}}\right)^{y_i}$, $y_i = 0, 1, \ldots$	μ_i	$\mu_i + \alpha\mu_i^2$
ZIP	$\begin{cases}\varphi_i+(1-\varphi_i)\exp(-\mu_i), & y_i=0\\ \frac{(1-\varphi_i)\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, & y_i=1,2,\ldots\end{cases}$	$(1-\varphi_i)\mu_i$	$(1-\varphi_i)(\mu_i+\varphi_i\mu_i^2)$
Hurdle	$\Pr[y_i=0/x_i]^{1-I_i}\times[1-\Pr[y_i=0/x_i]]\times\Pr[y_i/x_i, y_i>0]^{I_i}$. Where, $I_i = \begin{cases}1 & if \quad y_i\in\{1,2,\ldots\}\\ 0 & if \quad y_i=0\end{cases}$	$\Pr[y_i>0/x_i]\,E_{y_i>0}[y_i/y_i>0,x_i]$	$\Pr[y_i>0/x_i]V_{y_i>0}[y_i/y_i>0,x_i] + \Pr[y_i=0/x_i]E_{y_i>0}[y_i/y_i>0,x_i]$
Poisson Finite mixture	$\sum_{j=1}^{g}\frac{\pi_j\mu_j^{y_i}\exp(-\mu_j)}{y_i!}$	$\sum_{j=1}^{g}\pi_j\mu_{ij}$ Where: $\mu_{ij} = \exp(x_i'\beta_j)$	$E(V(y_i/z_i)) + V(E(y_i/z_i)) = E(y_i) + \upsilon_{ij}$ Where : $\upsilon_{ij} = \sum_{j=1}^{g}\pi_j\mu_{ij}^2 - \left(\sum_{j=1}^{g}\pi_j\mu_{ij}\right)^2$, $Z_i=(Z_{i1},\ldots,Z_{ig})$, $Z_{ij}=1$ when the subject i belongs to the j^{th} mixture component.

The socioeconomic data reported by each respondent includes information on sex, age, education and income. Education (EDU) is measured as the number of years in school and income (INC) is monthly household income/1000 in TND. Demographic variables are sex (SEX), indicated by a dummy variable (0 for male and 1 for female), and age (AGE) measured in years. The respiratory health situation is represented by four dummy variables, one for asthma (ASM), one for bronchitis or emphysema (BRO), one for hay fever or other forms of allergic runny nose (HAY) and one for other complaints from the nose: frequent sneezing, itch, blocked nose, etc. (OCN). The dummy variable receives a value of 1 if the respondent has, has had or has been diagnosed with asthma/bronchitis/hay-fever/other complaints from the nose; otherwise the value is 0. Data on smoking behavior are also included in view of the harmful effects of cigarettes on health. For this, we include the number of cigarettes smoked per day (SMK). Knowing that many individuals spent most of their time indoors, our study takes into account the effects of some characteristics of the respondent's house on his respiratory health status. Among the characteristics of the respondent's house that can affect his respiratory health, we include the number of persons (apart from himself) living permanently in his home (NBR). In addition, the home's date construction is included as an ordinal variable (HOM), which takes, respectively, the values 1, 2, 3, 4, 5 if the home was, respectively built after 1995, between 1980 and 1995, between 1960 and 1980, between 1940 and 1960 and before 1940. The outside temperature is also considered, for this the fourteen-day average of the daily maximum temperature is included (TEM). As a measure of air quality we use ozone measures in µg/m³ (OZN). Fixed effects in the form of type of zone dummies (urban (URB)/industrial (IND)/high traffic road (HIGH)/suburban (SUB)), are introduced to control for differences between the zones that are not accounted for by explanatory variables. For each respondent, the type of area in which he lives is indicated by the dummy that receives the value 1. When estimating the relationships between atmospheric pollution and lost work days, [15] emphasizes the importance of such variables in reducing the uncertainty attendant on the valuation of the effects of pollution on health.

Measures of the temperature are given by the National Institute of Meteorology. Air pollution data come from the National Environmental Protection Agency (NEPA). The NEPA was created in 1996 in order to analyze and monitor the state of the environment in the country. At the period of the study the network has 9 monitors of air pollution planted in different regions in Tunisia. Several studies have used more than one indicator of exposure to assess the effects of air pollution. A drawback of this is the high correlation between pollutants, both in time and space. To avoid multicollinearity, [16] recommended the use of only one pollutant as an indicator[5]. For the survey period, (March 2006-April 2006), the NEPA had monthly data on ozone measured in µg/m³, for the 9 stations of air quality measuring. March data on air quality were attributed to March respondents and the April data were attributed to April respondents.

The descriptive statistics show (see Table II) that the average number of RRADs during the 2-week recall period is 2.1 days. The average age is 35.137 with a 95% confidence level ranging from 34 to 36.2, and the proportion of women is

[5] To quantify the effects of air pollution, they used exposure-response functions for a 10 µg/m³ increase in particulate matter (PM_{10}). For the same reason [4] used only the NO_2 measure to identify CRRs in Sweden

44.7%. In average, the respondents smoked about 5.3 cigarettes a day, while the heaviest smoker reported smoking nearly 2 packs per day.

TABLE II
DESCRIPTIVE STATISTICS

Variable	Mean	95% CL	
Female (%)	44.7		
AGE, years	35.137	34.1	36.17
EDU, years of education completed	14.427	13.98	14.87
INC, monthly household income in 1000 TND	1.155	1.02	1.29
SMK, number of cigarettes smoked per day	5.268	4.36	6.18
ASM (%)	2.2		
BRO (%)	6.4		
HAY (%)	6.9		
OCN (%)	6.6		
HOM, (date of build) (%)			
Before 1940	3.4		
[1940 , 1960]	5.7		
[1960 , 1980]	22.1		
[1980 , 1995]	32.4		
After 1995	36.4		
NBR, number of people living with the respondent	3.231	3.05	3.41
OZN, µg/m³	69.098	64.31	73.89
TEM,C-degrees	17.982	17.9	18.07
Zone (%)			
Urban residence	33.2		
Industrial zone	12.5		
High traffic zone	34.9		
Sub-urban zone	19.4		
RRAD, days in the 2-week recall period	2.108	1.7	2.52
RRAD, binary (%)	27.8	23.45	32.15

As can be seen from Table II, the raw data are clearly overdispersed, since the mean number of RRAD is 2.1 with a variance equals to 17.9. Table III summarizes the distribution of RRADs for the whole sample of 407. As already announced, there is large number of persons (72%) reporting zero RRADs. A second mode corresponds to 14 RRADs.

Among the 113 reporting positive RRADs, 62 respondents have had a nose problem; 63 have had a throat problem and 36 have had a bronchial tubes problem. In addition to the large number of individuals reporting zero RRADs, our data present another feature showed in Table III, which is the large number of individuals reporting 7 and 14 RRADs and the absence of respondents reporting 8, 11 and 13 RRADs. This indicates the fact that some respondents may not accurately report the correct number of their RRADs. About the mode corresponding to 14 RRADs, it should be noted that individuals reporting 14 RRADs comprise those who have had respiratory problems for more than 14 days but were constrained to report the upper limit assumed by the response variable.

TABLE III
RRAD DISTRIBUTION

RRAD	No. OBS	%
0	294	72.24
1	2	0.49
2	9	2.21
3	17	4.18
4	21	5.16
5	4	0.98
6	2	0.49
7	18	4.42
8	0	0
9	1	0.25
10	2	0.49
11	0	0
12	1	0.25
13	0	0
14	36	8.84

IV. EMPIRICAL RESULTS

As announced below, we present the results for original sample and post-stratified samples. Results for the post-stratified sample are presented in Table IV. Results for the original are presented in Appendix 2.

TABLE IV
POISSON, NEGBIN, ZIP, POISSON HURDLE AND NEGBIN HURDLE ESTIMATION RESULTS

Variable	Poisson	Negbin	ZIP		Poisson hurdle		Negbin hurdle	
			Inflate	RRAD	Zeros	Positives	Zeros	Positives
Constant	15.492**	24.056***	-4.44	0.469	3.451	-0.014	5.572*	0.907
SEX	0.579***	1.060***	-1.197***	-0.254	0.901***	-0.219	1.172***	-0.274
AGE	- 0.004	- 0.007	-0.000	-0.003	-0.003	-0.001	-0.0005	-0.003
EDU	0.021	0.024	-0.094**	-0.036**	0.077**	-0.034**	0.091**	-0.038**
INC	- 0.131	- 0.414**	0.214	0.162*	-0.145	0.133	-0.205	0.167*
SMK	0.022**	0.039***	-0.046***	-0.009	0.036***	-0.008	0.045***	-0.011
ASM	0.723	1.316	-4.285	0.242	2.487	0.275	5.647	0.283
BRO	0.53	1.097***	-1.283**	0.075	0.938**	0.074	1.277**	0.119
HAY	0.835***	0.636**	-1.232***	0.323	0.948***	0.312	1.229***	0.323
OCN	0.691***	0.999**	-2.014***	-0.209	1.371***	-0.186	1.941***	-0.239
HOM	0.079	0.069	-0.053	0.149***	0.066	0.139***	0.06	0.157***
NBR	0.094**	0.094	-0.180***	0.037	0.142**	0.037	0.178**	0.049
OZN	0.022**	0.037***	-0.026***	-0.005	0.019***	-0.005	0.027***	-0.004
TEM	- 0.955**	-1.489***	0.570***	0.106	-0.464***	0.13	-0.633***	0.074
URB	- 0.980**	-1.444***	-	0.094	-	0.119	-	0.185
IND	- 0.616	-1.520***	-	0.484	-	0.482	-	0.481
SUB	- 0.387	-0.343	-	0.336	0	0.361	-	0.448
LL	-1125.22	-604.933	-	-822.813	-189.063	-352.643	-187.811	-329.44
AIC	2267.434	1226.866	-	1676.626	-	1114.512	-	1065.502
α	-	4.677***	-	-	-	-	0.161***	-

Note: ***, **, *: significance at the 1%, 5% and 10% level.

The post-stratification has the advantage of sample bias correction. It is undertaken here by drawing with replacement 200 samples of size 407 from our original sample according to the population frequencies given in Table VII see Appendix 2). On the basis these samples[6], we construct confidence intervals for the mean and the variance of RRADs. The mean and variance confidence intervals at 1% level risk are respectively [1.828, 2.783] and [13.916, 22.571], this indicates that a strong overdispersion is also characterizing the post-stratified samples.

According to the AIC and the log-likelihood ratio tests, the Negbin hurdle model overcomes all the alternative models. Hence, this model better fits the data. For example, the LR test statistics of Negbin against the hurdle Poisson is equal to 126.454 (p-value=1%), the LR statistics of Poisson hurdle against the Negbin hurdle is equal to 48.909 (p-value=1%).

Negbin hurdle results show that women, smokers, those who have had respiratory problems and those with more people living permanently in their homes are more likely to suffer from respiratory illnesses. OZN is also significant in determining the incidence of such days. Temperature is likely to reduce the incidence of RRADs.

In the conditional mean part, only HOM is significant to determine the duration of RRADs with positive sign, indicating that those living in older homes have more RRADs. The variables INC and EDU are also significant but the sign of INC is not as predicted.

In order to take in account of any unobserved heterogeneity in the population groups, we estimate FM models. To determine the number of components in the mixture we use the approach of [13] presented above. Our saturated models involve the following covariates: Constant, SEX, AGE, EDU, INC, SMK, DIS, HOM, NBR, OZN, and TEM. Some covariates are dropped to avoid over-parameterization. For the same goal, we represent the health situation of each respondent by the variable DIS which receives a value of 1 if the respondent has, has had or has been diagnosed with one of the four health indicators introduced in the preceding models; otherwise the value is 0. Maximum likelihood estimates are calculated by using the EM algorithm. For the choice of starting values, we divide the data into g components (for g=2, and g=3)[7] according to the information given by the indicator variable $Z_i = \left(Z_{i1},\ldots,Z_{ig}\right),\ i = 1,\ldots,407$.

A multiple linear regression model is then performed for each component. Least squares estimators of the parameters are then used as starting values for each component in the Poisson mixture. The estimates of the parameters in this model are then used as starting values. Starting values for the logit model are obtained in the same way. Estimation of saturated models is given in Table V. In the 2-components finite mixture (FM2), we suppose a decomposition of the population on two groups, the first is the group of those who have had a number of RRAD lower than 4.

The second group is formed by the remainder individuals who have had a number of RRAD equal or upper than 4. In the 3-components finite mixture we suppose a decomposition of the population into three groups: the first is formed by those who have had less than 4 RRADs, the second by those who have had a number of RRADs that is equal or greater than 4 and less than 12 and the third group is formed by the remainder. The BIC is minimized with the two components model which fits best the data than the FM with three components.

The results relative to the two components model show that the variables SEX, EDU, SMK, OZN are significant with expected signs in predicting the Poisson rate for the first group, whereas only the constant and HOM are significant in predicting the Poisson rate for the second group. In modeling the mixing proportions π_j, the variables SEX, DIS, NUM, OZN and TEM were significant with predicted signs.

To test the FM2 against the Negbin hurdle alternative, the calculated value of the LR test statistic is -2(-551.091+517.251)=67.679. The 1% critical value is 12.383, and then the FM2 is rejected against the Negbin hurdle.

If g=3 is preferred, then we test between g=3 and g=4 components.

The interpretation of regression coefficients, in particular the OZN coefficient is of great interest. Hence, it allows us to illustrate the changes in the number of RRADs that might be associated with changes in ozone levels, known as semi-elasticity. For the hurdle Negbin and the finite mixture specifications, semi-elasticities are evaluated as indicated in (3) to (8):

$$\frac{\partial E\left[y/x\right]}{\partial x_j} = \frac{\partial E_{y>0}\left[y/y>0,x\right]}{\partial x_j}\Pr\left[y>0/x\right] + E_{y>0}\left[y/y>0,x\right]\frac{\partial \Pr\left[y>0/x\right]}{\partial x_j}. \quad (3)$$

Obviously y stands for the number of RRADs that an individual have had during the last fortnight. For the estimation results of the hurdle specifications (Table IV), the OZN was significant in the determination of the incidence of respiratory problems but not in the determination of their duration. Then the first term of (3) is set equal to zero. From the same equation, the change in the number of RRADs that might be associated with a change in the OZN level is calculated as follows:

$$\frac{\partial E\left[y/x\right]}{\partial OZN} = \frac{\beta_{OZN}e^{x'\beta}E\left[y/y>0,x\right]}{\left(1+e^{x'\beta}\right)^2}. \quad (4)$$

[6] These samples will then be used for all econometrical estimations. Confidence intervals for the parameters estimated are based on the "bootstrap-t" approach. We use B′=2000 for confidence intervals construction.

TABLE V
FINITE MIXTURE MODELS FOR POST-STRATIFIED SAMPLE

Variable	2 components Finite mixture			3 components Finite Mixture				
	Comp 1	Comp 2	logit (p)	Comp 1	Comp 2	Comp 3	logit (p)	logit (p)
Constant	11.42	7.79**	-7.70*	13.86	9.05	-4.28	-19.79	-15.75
SEX	1.95***	-0.03	-0.97***	-0.03	-0.37**	0.16	-1.65**	-0.66
AGE	-0.02	-4.10⁻³	2.10⁻³	-0.03	3.10⁻⁴	4.10⁻³	7.10⁻³	6.10⁻³
EDU	0.45***	-4.10⁻³	-10⁻³	0.76	-0.05***	-4.10⁻³	-0.10	-0.11
INC	-1.26	0.01	0.02	-1.97	0.10	0.09	0.28	0.28
SMK	0.12***	-3.10⁻⁴	-0.01	0.21	-4.10⁻²	2.10⁻³	-0.04	-0.05
DIS	1.67	0.18	-1.32***	2.92	-0.33*	0.05	-2.54***	-1.99***
HOM	0.43	0.09*	0.04	0.64	-0.08	-0.01	-0.27	-0.40
NBR	0.21	0.03	-0.19**	0.03	-0.04	0.03	-0.34*	-0.20
OZN	0.04***	3.10⁻²	-0.02***	0.06	3.10⁻³	-6.10⁻³	-0.04**	-0.01
TEM	-1.47*	-0.33	0.67***	-2.08	-0.35	0.38	1.65*	1.20
	LL=-551.091, m=33, BIC=1300.473			LL=-495.876, m=55, BIC=1322.236				

For the FM2 specification the mean is calculated in the following way (1 and 2 denote the first and second component of the mixture):

$$E[y] = E[y/1]P(1) + E[y/2]P(2) = \exp(z'\alpha_1)\frac{\exp(x'\beta)}{1+\exp(x'\beta)} + \exp(z'\alpha_2)\frac{1}{1+\exp(x'\beta)}. \quad (5)$$

Denoting

$$\mu_1 = \exp(z'\alpha_1), \mu_2 = \exp(z'\alpha_2), \pi = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$$

and $1-\pi = \frac{1}{1+\exp(x'\beta)}$ gives,

$$\frac{\partial E[y/x]}{\partial OZN} = \alpha_{1OZN}\mu_1\pi + \mu_1\frac{\partial \pi}{\partial OZN} + \alpha_{2OZN}\mu_2(1-\pi) - \mu_2\frac{\partial \pi}{\partial OZN}. \quad (6)$$

with,

$$\frac{\partial \pi}{\partial OZN} = \frac{\beta_{OZN}\exp(x'\beta)}{(1+\exp(x'\beta))^2} = \beta_{OZN}\pi(1-\pi). \quad (7)$$

Then,

$$\frac{\partial E[y/x]}{\partial OZN} = \mu_1\pi(\alpha_{1OZN} + \beta_{OZN}(1-\pi)) + \mu_2(1-\pi)(\alpha_{2OZN} - \beta_{OZN}\pi). \quad (8)$$

The changes in acute respiratory health that might be associated with changes in ozone levels are given in Table VI, which gives an estimation of the proportionate change in the conditional RRADs mean if the OZN changes by one unit according to different specifications.

For the Negbin hurdle and the FM2 specifications, the intermediate calculation is given in Table VI, giving rise to 0.020 and 0.005 as respective OZN semi-elasticity values. These values will able us to monetize the ozone effect on human health.

TABLE VI
OZONE SEMI-ELASTICITIES

	Specification	
	Negbin hurdle (c=0)	FM2 (c=3)
E[RRAD|x]	1.951	1.639
E[RRAD|RRAD>c]	7.656	8.404
Pr[RRAD>c]	0.255	0.193
OZN semielasticity	0.020	0.005

In the literature, the value of air pollution impacts depends on the value assigned to human health. Studies of the value of human health simply reflect what people seem willing to pay in monetary units for a marginal change in health. To evaluate the effect of a change in the level of ozone, suppose that a policy measure reduces by 1 unit the level of ozone concentration. Using the ozone semi-elasticity, this measure will imply a reduction in the expected number of RRADs which ranged between 0.5% and 2% depending on the model specification used.

Given the RRAD mean calculated on the basis of post-stratified samples which is 2.295 days, this policy measures will simply an average individual reduction that ranged between 0.011 and 0.046 day each two weeks.

Finally, a monetary value could be assigned to this hypothetical improvement in health. To this end, the contingent valuation method is used. The willingness to pay to avoid one day of respiratory restricted activity was elicited by each respondent. The mean value elicited was 0.963 dinar per day. Therefore, the value of the average individual reduction per 2 weeks is ranged between 0.01 and 0.044 dinars. That is, the annual benefit per person of the 1 unit reduction in the ozone concentration will be valued at as much as 1.152 dinar. Finally, the annual benefit to the urban population is estimated at 7.2 million dinars.

V. DISCUSSION

Following [1]-[4], we estimated several models adjusting RRAD and socio-economic and air quality measures. The models considered are Poisson, Negbin, ZIP, Hurdle Poisson, Hurdle Negbin and Finite mixture. The Hurdle Negbin performs all the other models. We found encouraging results in accordance with previous studies finding. Indeed, [2] has used three indicators of morbidity measured for a 2-week

period, among which the number of RRADs. He used Poisson specifications for data provided by the annual health interview surveys (HIS) for 1976-1981. Pollution is measured by the 2-week average of fine particles which received 0.7 to 2.2 percent as coefficient. In [3] and [1] the number of RRADs is also considered as dependent variable. The estimation samples come from the 1979 HIS in the United States. For [3], Poisson specifications were used. The mean of individual elasticity obtained is ranged from 0.006 and 0.485. In [1] linear estimators have been used. The estimated elasticity for ozone ranged from 0.12 to 2.97. Reference [4] estimated a CRR based on Swedish national health survey data. To deal with overdispersion characterizing the number of RRADs they used a logit model for estimating probability of RRAD > 0. The individuals with positive RRADs are analyzed separately using a Poisson model. Their results indicate that if the level of NO_2 increases by one unit ($\mu g/m^3$), the number of RRADs will increase by 3.2 percent.

The difference between these results is mainly due to the specificity of each country, in addition to their dependence on the estimation methodology. Compared to these results, ours can be improved if a larger sample were available.

VI. Conclusion

The morbidity effects of air pollution in Tunisia are investigated. The data were elicited using a questionnaire surveying individuals from 9 regions in Tunisia for which air pollution is monitored. In order to improve the quality of our data, post-stratification and bootstrapping techniques were used.

The dependent variable is the number of RRADs experienced by each respondent during the 2-week period prior to the date of the interview. This variable is characterized by the simultaneous presence of excess zeros and an overdispersion detected by several tests. To handle with this characteristic, models allowing for overdispersion and zero-inflation were estimated.

In terms of goodness of fit, measured by either log-likelihood or AIC, hurdle Negbin is the best to fit our data followed by Poisson hurdle, FM2, Negbin, ZIP, and Poisson, in that order. The superiority of the hurdle Negbin could be explained by the joint presence of unobserved heterogeneity and a considerable excess of zeros. Beyond the estimation method, our results indicate that women, smokers, those with more people living permanently with them are found to be more likely to suffer from respiratory illness. The results indicate notably that the population is sensitive to air pollution: increases in modest pollution levels tend notably to determine the incidence of respiratory problems and to also prolong their duration (depending on the model specification). It was found that a reduction of one unit in the ozone level leads on average to a decrease ranged between 0.5% and 3.7% in the expected number of RRADs depending on the specification used. This corresponds to an average individual reduction that ranged between 0.011 and 0.046 day each two weeks.

The contingent valuation method is used to monetize this hypothetical improvement in health. The willingness to pay to avoid one day of respiratory restricted activity was elicited by each respondent. The mean value elicited was 0.963 dinar per day. Therefore the value of the average individual reduction per 2 weeks is ranged between 0.01 and 0.044 dinars.

The annual benefit for each individual of this reduction will be valued at as much as 1.152 dinar. Therefore, the annual benefit to the adult urban population of Tunisia is valued at as much as 7.2 million dinars.

Appendix

Appendix 1: Post-Stratification and Bootstrapping

To implement the Post-stratification technique, the population is classified by three criteria: sex (1 female, 2 male), 4 categories of age (1 age 18 to 29, 2 age 30 to 39, 3 age 40 to 49, 4 age 50 to 60), and 9 zones (1 Bab Sâadoun, 2 Nahli, 3 El Mourouj, 4 Ben Arous, 5 Rades, 6 Bizerte, 7 Sousse, 8 Sfax Siap, 9 Sfax). This classification implies a division of the population into 2×4×9=72 post-stratification cells.

The size of each cell according to the parent population is derived from the 2004 Tunisia Census of Population, which provides the population structure by age group, sex and delegation. Then we derive a classification of 407 individuals according to the parent population composition of size N=1 603 906. Table VII (respectively Table VIII) gives the size of each cell according to the parent population (respectively to our sample).

TABLE VII
Population Classification of 407 Individuals

Sex	Male				female			
Zone/Age	1	2	3	4	1	2	3	4
1	26	17	15	10	25	17	15	10
2	8	6	4	3	8	5	4	3
3	5	4	4	2	5	4	3	1
4	3	2	2	1	3	2	2	1
5	6	5	4	3	6	5	4	3
6	5	3	3	2	5	4	3	2
7	10	6	5	3	9	6	5	3
8	11	7	7	4	10	7	6	4
9	5	3	2	1	4	3	2	1

TABLE VIII
Sample Classification of 407 Individuals

Sex	Male				female			
Zone/Age	1	2	3	4	1	2	3	4
1	9	11	6	10	27	5	14	9
2	10	11	2	1	6	12	2	3
3	13	6	1	4	6	1	1	1
4	2	2	3	2	4	2	1	2
5	15	8	10	2	16	9	7	2
6	11	16	5	2	10	2	4	2
7	2	1	2	1	3	3	1	1
8	13	13	7	7	5	1	3	1
9	6	3	5	3	3	4	8	1

In order to test for sample representativeness, we compare between the sample and the population frequencies according to the variables used as criteria for the post-stratification (sex, age and zone). To this end, the Khi test is used for testing the null hypothesis: the population and the sample come from the same distribution. For confidence intervals construction, we use the approach of percentile confidence interval based on percentiles of the bootstrap distribution of a statistic as well as the bias corrected version of this interval as presented in [17]. Our calculations are obtained with 2000 samples of size n=407 from our original sample. The percentile confidence interval at α=1% risk level is [44.557; 109.382]. To improve these results, we use the bias-corrected and accelerated approach (see [17]). The interval is almost the same as the 99% BC_a interval with $[\alpha_1,\alpha_2]=[0.726, 99.662]$ which is [45.941, 111.660]. The calculated value obtained on the base of the original sample is 259.679, these results allows us to reject the null hypothesis that population and the sample come from the same distribution. The rejection of the null hypothesis implies that frequency distribution of the sample at hand, at least according to stratification criteria such as age, sex and region, is quite different from what is observed at the population level. That is any econometric result based on the original sample without any treatment permitting to overcome this shortcoming will lead definitely to both biasness and inconsistency of the estimates, which preclude understanding the true interconnection between the key variables. In order to fix this problem the idea is to use post-stratification techniques based on population frequencies to draw consistent poststatificated samples using bootstrapping. Our econometric results are based on independent bootstrap samples.

Appendix 2

TABLE IX
POISSON, NEGBIN, ZIP, POISSON HURDLE AND NEGBIN HURDLE ESTIMATION RESULTS FOR THE ORIGINAL SAMPLE

Variable	**Poisson**	**Negbin**	**ZIP**		**Poisson hurdle**		**Negbin hurdle**	
			Inflate	RRAD	Zeros	Positives	Zeros	Positives
Constant	13.175***	12.683	-0.21	1.371	0.719	1.245	0.305	2.947
SEX	0.629***	1.086***	-0.937***	0.091	0.802***	0.091	0.935***	0.091
AGE	-0.003	-0.006	0.011	0.006*	-0.012	0.007*	-0.011	0.006
EDU	0.019**	0.008	-0.049	-0.008	0.048**	-0.007	0.049	0.011
INC	-0.267***	-0.588***	0.402*	0.085	-0.316*	0.085	-0.399*	0.094
SMK	0.024***	0.051***	-0.049***	-0.004	0.041***	-0.004	0.049***	-0.005
ASM	0.811***	1.327	-1.289	0.322**	0.583	0.323**	1.284	0.32
BRO	0.432***	0.843	-1.212**	-0.046	0.813**	-0.047	1.277**	0.006
HAY	0.572***	0.542	-1.228***	-0.01	0.936***	-0.009	1.227***	-0.03
OCN	0.842***	1.215**	-2.035***	-0.104	1.449***	-0.103	2.026***	-0.107
HOM	0.069**	0.178	-0.105	0.034	0.093	0.037	0.105	0.053
NBR	0.117***	0.185**	-0.198***	0.059**	0.143***	0.061**	0.198***	0.066
OZN	0.020***	0.028**	-0.016**	0.0005	0.014**	0.0005	0.016**	0.003
TEM	-0.824***	-0.85	0.239	-0.0007	-0.248	0.004	-0.245	-0.105
URB	-0.586***	-0.734	-	0.094	-	0.096	-	0.143
IND	-0.435**	-1.171*	-	0.24	-	0.244	-	0.111
SUB	-0.217	-0.047	-	0.282*	-	0.284*	-	0.368
LL	-1164.976	-570.807	-	-712.992	-186.342	-360.886	-187.811	-318.053
AIC	2346.952	1158.614	-	1456.984	-	547.228	-	536.864
α	-	5.672***					0.279***	-

TABLE X
FINITE MIXTURE MODELS FOR THE ORIGINAL SAMPLE

	2 components Finite mixture			3 components Finite Mixture				
Variable	Comp 1	Comp 2	logit (p)	Comp 1	Comp 2	Comp 3	logit (p)	logit (p)
Constant	11.45	5.21**	-0.47*	6.61	7.98**	0.43	-6.24	-7.98
Sex	1.06***	0.06	-0.75**	1.60**	-0.06	0.03	-1.07**	-0.39
Age	-0.05***	-0.003	-0.002	-0.09**	-0.001	0.000	-0.001	-0.009
Edu	0.11***	-0.008	-0.027	0.25***	-0.02	-0.004	-0.03	-0.02
Inc	-1.04***	0.02	0.30	-0.62*	0.05	0.02	0.20	-0.26
Smk	0.09***	0.001	0.028*	0.14***	0.005	0.001	-0.03	-0.008
Dis	2.99***	0.19**	-1.09***	2.86***	-0.23*	-0.001	-1.86***	-0.635
Hom	-0.07	0.016	-0.10	-0.28	-0.049	0.011	-0.14	-0.01
Num	0.19**	0.03	-0.19***	0.13	-0.03	0.01	-0.30***	-0.13
Ozn	0.03***	0.003	-0.015**	0.03*	0.005	-0.002	-0.02**	-0.01
Tem	-0.97*	-0.19	0.259	-0.82	-0.35	0.13	0.72	0.63
	LL=-556.6747, m=33, BIC=1311.640			LL=-506.6057, m=55, BIC=1343.696				

ACKNOWLEDGMENT

We acknowledge the National Environmental Protection Agency for providing ozone data for this study.

REFERENCES

[1] J. Mullahy, P. Portney, "Air pollution, cigarette smoking, and the production of respiratory health", J Health Econ 9, 1990, pp. 193-205.
[2] B.D. Ostro, "Air pollution and morbidity revisited: A specification test". J Environ Econ Manage 14, 1987, pp. 87-98.
[3] P. Portney, J. Mullahy, Urban air quality and acute respiratory illness. Resources for the future/Washington, D.C., 1986.
[4] E. Samakovlis, A. Huhtala, T. Bellander, et al. "Air quality and morbidity: Concentration-response relationships for Sweden". NIER Stockholm Working Paper No 87, 2004.
[5] J. Mullahy, Heterogeneity, "Excess Zeros, and the Structure of Count Data Models", J Appl Econometr 12, 1997, pp. 337-350.
[6] J. Mullahy, "Specification and testing of some modified count data models". J Appl Econometr 12, 1986, pp. 337-350.
[7] M.L. Dalrymple, I.L. Hudson, RPK Ford, "Finite Mixture, Zero-inflated Poisson and Hurdle models with application to SIDS", Comput Stat Data Anal 41, 2003, pp. 491-504.
[8] S. Gurmu and P. Trivedi, "Excess Zeros in count models for recreational trips", J Bus Econom Statist 14 No 4, 1996, pp. 469-477.
[9] A.C. Cameron, P.K. Trivedi, Regression analysis of count data. Cambridge University Press. 1998.
[10] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", J R Statist Soc B 39 No1, 1977, pp. 1-38.
[11] G. McLachlan and D. Peel, Finite mixture models. A Wiley-Interscience Publication, 2000.
[12] A.J. Izenman and C.J. Sommer, "Philatelic mixtures and multimodal densities", J Amer Statist Assoc 83, 1988, pp. 941-953.
[13] P. Wang, M.L. Puterman, I. Cockburn et al. "Mixed Poisson models with covariate dependent rates", Biometrics 52 No2, 1996, pp. 381-400.
[14] D. Holt D, TMF Smith, "Post stratification", J R Statist Soc A 142 No1, 1979, pp. 33-46.
[15] J. Hausman, B. Hall, Z. Griliches, "Econometric models for count data with an application to the patent R&D relationship", Econometrica 52 No 4, 1984, pp. 909-938.
[16] N. Künzli, R. Kaiser, S. Medina et al. "Public health impact of outdoor and traffic-related air pollution: A European assessment" The Lancet 356, 2000, pp. 795-801.
[17] B. Efron and R. Tibshirani, An introduction to the bootstrap. Chapman & Hall. 1993.

Mokhtar Kouki is Professor at the Higher School of Statistics and Data Information, Tunis, Tunisia (from September 2003). October 2011-December 2014: Senior Econometrician creativ-ceutical. HEOR (Luxembourg). 1997-Now: Senior researcher at Tunisia Polytechnic School.

Inès Rekik is PhD, assistant professor at the Higher Institute of Computer Science of El Manar. Tunisia.