REVIEW PAPER

International Journal of Occupational Medicine and Environmental Health 2023;36(6):704–716
https://doi.org/10.13075/ijomeh.1896.02237

TIME SERIES ANALYSIS IN ENVIRONMENTAL EPIDEMIOLOGY: CHALLENGES AND CONSIDERATIONS

SANDRA GUDZIUNAITE[1], ZANA SHABANI[2], LISBETH WEITENSFELDER[1], and HANNS MOSHAMMER[1]

[1] Medical University of Vienna, Vienna, Austria
Department of Environmental Health, Center for Public Health
[2] University of Hasan Pristina, Pristina, Kosovo
Medical Faculty

Abstract
In environmental epidemiology, time series analyses represent a widely used statistical tool. However, though being commonly used, there is soften confusion regarding the specific requirements, such as which link function might be most appropriate, when or how to control for seasonality or how to account for lags. The present overview draws from experiences in other disciplines and discusses the proper execution of time series analyses based on considerations that are relevant in environmental epidemiology. Time series analysis in environmental epidemiology focuses on acute events caused by short-term changes in exposure. These exposures should be fairly wide-spread affecting a large number of persons, usually all inhabitants of a political entity. Pollutants in air or drinking water as well as meteorological factors serve as typical examples. Despite the many time series analyses performed world-wide, some health effects that would lend themselves to that approach are still under-explored. This would include also some neurological and psychiatric endpoints. Int J Occup Med Environ Health. 2023;36(6):704–16

Key words:
statistical methods, regression models, environmental epidemiology, short-term effects, time series analyses, confounder control

INTRODUCTION

Time series analysis (TSA) can be broadly described as the study of time series data, which is characterized by having been sampled or clustered at equally spaced intervals. Common examples include the effect estimates per year [1], sales per month, births or conceptions per week [2,3], suicides per day [4,5], or heart beats per minute. Time series analysis is used in a wide range of fields including physiology [6], economics [7] (e.g., by studying the correlation between economic growth and electricity consumption) [8], ecology (e.g., for predicting the migratory patterns of fish) [9] and genetics (e.g., by exploring biological processes through gene expression) [10]. Each field of research can investigate different aspects of time, and will therefore adopt a method that better suits their intents.

A properly executed time series analysis can, in fact, be used for forecasting [11], regression modelling [12],

Funding: this research was supported by the Austrian Development Cooperation and the Government of Kosovo (grant No. K-02-2021, within the funding scheme HERAS plus [Higher Education, Research and Applied Science plus] for the project entitled "Air Pollution in Kosovo – impact on human health, behaviour change and policy recommendations," project manager: Hanns Moshammer).
Received: June 2, 2023. Accepted: August 24, 2023.
Corresponding author: Hanns Moshammer, Medical University of Vienna, Department of Environmental Health, Center for Public Health, Kinderspitalgasse 15, 1090 Vienna, Austria (e-mail: hanns.moshammer@meduniwien.ac.at).

or trend and seasonality extraction [13]. Since time can be investigated for several reasons, the methods used need to reflect the study's aim. Each field has somewhat contributed to the rich literature of methodologies employable for time-series studies, that with time have interacted with each other and have resulted in a rich toolkit available to approach the issue. It follows that, as methods are developed in different fields, and as software and programs are created to meet their demand, each one of them will be shaped by the aims that drove its initial formulation.

Though often generally applied in environmental epidemiology, there is a broad variety of research questions that use TSA as a tool. However, different applications require more or less different statistical methods which might lead to confusion regarding requirements or correct implementations of the method. Hence the present methodological discussion outlines types of research questions and their possible TSA applications in the field of environmental epidemiology.

Indeed, this overview was originally born out of an intention to emulate study results from another study (regarding air pollution and pregnancy loss [2]) with own data. Before the authors analyzed the weekly birth data from Vienna in relation to air quality during pregnancy, they aimed to understand sources of autocorrelation and possibly overdispersion that would – if not controlled for correctly – eventually bias their later results [3]. The use of weekly data in TSA is rare and therefore experience regarding the appropriate way to control for longer-term temporal variations (e.g., seasonal variation) is poor in that case.

In the environmental health context, TSAs have often been used to analyze the acute development of morbidity or mortality in dependence of short-term variation of environmental factors. Examples include mortality and exposure to various forms of air pollution [14,15], or the increasingly more relevant issue of the acute effects of heat-waves [16]. Chronic diseases and endpoints that develop slowly or with a long latency period like cancer or many neurological and psychiatric diseases, lend themselves rather not to TSA. But as an exception of that rule, suicides [4,5,17] and cerebral insults [18–22] have been successfully analyzed in TSA. In environmental epidemiology, this method has been one of the standard approaches to assess impacts of environmental factors on acute non-infectious diseases like cardiovascular deaths [15], hospital admissions because of respiratory diseases [23] or GP consultations [24,25], with conventionally generalized linear models (GLMs) or generalized additive models (GAMs). However, the same analysis practices are often observed with infectious diseases despite of the substantial differences from non-infectious diseases that may result in analytical challenges [26].

METHODS

This paper offers an overview of how the issue of time has been dealt with in the field of environmental epidemiology. Its goal is to provide hints for both statistical applicants and readers of time series analyses on what to consider for interpretation and which pitfalls to avoid in environmental epidemiology. The authors have been motivated to reflect on TSA and its methodologies after inferring that the methods developed to explore time series type of data were quite diverse, which meant that failing to pinpoint the aims and assumptions behind the statistical tests carried the potential to result in confusion and, possibly, misunderstanding.

To that end, the authors performed a non-systematic literature review with a focus on preferably newer papers applying time series analyses, not only but mostly from the environmental epidemiology field. The authors' aim was not to provide a complete picture of this interesting instrument, but rather demonstrate the breadth but also the difficulties that one encounters when investigating time series. Hence for reaching this goal, an all-encompassing review was not necessary, and the literature

search ended when the authors found examples for each of the below mentioned considerations.

RESULTS

Choosing the appropriate link function

When analyzing time series, one can isolate the trend and seasonal component in various ways:

1) modelling by GAM or GLM, which follows broadly the methods described by Bhaskaran et al. [12],
2) adaptations to the autoregressive integrated moving average (ARIMA) model, which originate from econometrics, and proved to be particularly useful for studying the trend and seasonal patterns of a given time series, a technique that has been adapted in the past decade into seasonal ARIMA with exogenous variables (SARIMAX) [27].

The following pages will attempt to summarize the insights gained from looking at the methodologies adopted in recent time series studies in environmental epidemiology, published in recent years, and reflect on the challenges associated with translating the most popular methods to this field.

A question that arises automatically is how to link the response variables to the relevant independent variables.

To conclude which one should be the most appropriate link function, one should consider some unique properties of time series data. A time series is usually a sequence of data points collected and aggregated at equally spaced points in time. In environmental epidemiology these values often indicate the daily, weekly or monthly numbers of events like deaths, births or hospital admissions. Therefore, as a rule of thumb, the independent variable in a time series analysis in environmental epidemiology is usually a count value. The dataset in which they are compiled is restricted to taking positive integer values. This condition calls for a Poisson regression model in case the model's assumptions are met. The Poisson distribution expresses the probability of a given event (x) occurring at a set time (t) given that the average number of its occurrences (λ) is equal to its variation [28].

There are instances in which the choice of a Poisson model might be reconsidered. The Poisson model asks for some specific assumptions that are not met by all count datasets. In a Poisson regression, the dataset should be equidispersed, meaning that its variance is equal to its mean. Most commonly, we will find ourselves in front of a case of overdispersion. This roadbump is relatively common, and if unaccounted for, the resulting time series model will be misleading, with an underestimation of the parametric standard deviation and an overestimation of its significance value [29].

Overdispersion is usually accounted for by changing the model linking function from Poisson to quasi-Poisson, or to a negative binomial [30]. Examples of overdispersion being accounted for using a quasi-Poisson distribution are offered by Gu et al. [31]. Negative binomial distribution has been used by Nayebare et al. [32]

Todkill et al. [33] offer an example of how to choose between the 2 model specifications. They selected the model with the lowest dispersion parameter. Both methods can be easily implemented in most statistical software, like STATA and R.

Seasonality

Overdispersion is only one of the potential issues to be encountered whilst modelling time series data. A further assumption of a Poisson model that is not always met in a raw data set is the assumption that the measurements are independent from each other. For this reason, one does not only input the time series data into the statistical software and call it a "model". Time series data is inherently "dirty", and requires care and polishing before it can indicate meaningful results [7].

As aforementioned, TSA is best suited to investigate the short-term effects of an effector on the population. Therefore, a time series in its raw form is likely to be strongly affected by its seasonal component.

This issue is statistically demonstrated by the fact that the raw time series datasets present high autocorrela-

tion rates. Autocorrelation is defined as the correlation between the measured variable at any given time-point (t) and at several time points (t-y) shortly before. For example, a "typical" environmental raw data set might include temperature as independent variable and daily death counts as dependent. But daily deaths as well as temperature do follow a seasonal pattern that needs to be considered.

In epidemiological reasoning, the idea of seasonality in a dataset might be intuitive given that the number of deaths, births or the disease onset might be heavily influenced by other, external factors. As an example, the number of daily deaths, which has long been established to be influenced by temperature, can be expected to be higher in the winter than in the summer. A scatter plot of values over time usually is enough to display that the time series exhibits a clear seasonal trend.

If this trend is unaccounted for, the short-term effects of exposure are undetectable behind the louder noise of the seasonal pattern of the response variable. This is a problem common to most time series, which means that a relatively wide number of solutions have been suggested.

A very simple, and somewhat rudimentary way for accounting for seasonal variations is to disregard them continuously and divide the time series into temporal categories instead. This principle has been described as the time stratified model [12] and it can be achieved by cutting snippets of regular time frames, such as month, season and so on.

The reasoning behind it is that the recurrent time frame of a series can be split into its components, taking into account that each component is expected to be exposed to similar exogenous factors, and compare them across recurrences of the different exposure cycles.

A recent example of this method being used is offered by Nayebare et al. [32] who studied effects of fine particle exposure on cardiopulmonary morbidity in Jeddah. They sampled their measurements for 6 weeks for each yearly quarter, each quarter representing a sampling cycle, to then insert the measurements into a GLM linked with binomial distribution.

Its main criticism is a reflection of its strength: it is a simple approach. As such it does tend to overlook the overall changes in trend of exposure and response variables throughout the time series, and in turn, from one time stratum to the other.

Some have exploited the apparently recursive nature of the seasonal component of a time series, and fit them to sinusoidal wave functions. The reasoning at the core of this method is that the sinusoidal waves filter out from the regression all patterns of a given wavelength, by providing the model with a term that reflects a full seasonal cycle. Examples of such solution are exemplified in Moshammer et al. [34]. This method is also rather old and hence also has been criticized already by Schwartz et al. [28] in 1996 as rather inflexible ("One concern with such a model is that it assumes that the seasonal peak is the same height and occurs at the same time each year. However, there are patterns in the intensity of, for example, influenza epidemics, with 2 year cycles both observed in practice and justified by mathematical modelling."). This method makes the strong assumption that the seasonal component to be captured is very regular in frequency and amplitude, and that it repeats itself invariably throughout the series. But especially when the focus of the study is on astronomical factors like daily duration of sunshine [35], sinusoidal wave functions seem the best choice. And in the end, also seasonal variations of meteorological variables over a span of some years are driven by astronomical influences mostly.

A more sophisticated way to go around the problem is by using a Fourier series. A Fourier series is a harmonically related, weighted series of sine waves (inherently periodic). Including multiple waves makes up for a more flexible, resulting function [12]. Thus the frequency for

which such waves are incorporated in the model could be determined by statistical considerations.

Examples of recent studies using this method for seasonal adjustment are offered by Todkill et al. [32], who modelled seasonal trends using Fourier terms with 4 degrees of freedom per year integrated in a distributed lag nonlinear model (DLNM), and Aik et al. [36] who controlled for seasonal variation by using Fourier terms to de-seasoning their climatic variables before including them into the regression analysis, and included a pair of periodic functions to control for seasonality of their reports.

Amongst the criticisms that stick out for this method, the most prominent one regards the assumption that the frequency and amplitude of the seasonal component are constant. This is both a weakness and a strength, as from one side it allows for a simple account for seasonality, but from the other, it does not provide the most accurate results, as the wavelength and amplitude of the sinusoid in environmental data can vary. This can render it an unsuitable method for data illustrating irregular phenomena, or those whose outcomes depend on exogenous stressors (e.g., climate).

Although earlier papers mention a wide range of smooths to control for seasonality (moving averages, kernel smoothing) most recent literature approaches this issue using splines and locally weighted scatterplot smoothing (LOWESS or LOESS)[37].

Moving averages and weighted moving averages are rudimentary ways of determining what are the expected number of cases of the value of interest in a "business as usual scenario", whilst unaffected by short term effects of the variables of interest and confounders varying on a similar narrow time-scale. This method assumes that the leftover noise fluctuations from the expected mean can be explained by short term effects of exogenous values.

The "smoothness" of the function is given by the size of the smoother span. It results in a (usually either linear or quadratic) curve that has been generated by fitting least squares for a proportion of the data predefined by the smoother span, weighted depending on the distance of the data points within the smoother span. It can be implemented in R with the loess function through the package "modreg". Typical smoother spans in a set of daily data in order to control efficiently against seasonal variation would likely take the size of a month.

A popular alternative for controlling for trend and season is to use flexible spline functions. A spline is a mathematical representation of a flexible shape. This method allows for the end-to-end conjunction of low-order polynomial functions capturing the fluctuations of the time series curve [38]. The structure of a spline is customizable, and the user has control over how "wavey" it is. To ensure that the ends of the composing functions meet smoothly at their junctions (or "knots"), one should recur to restricted or natural splines, in which the derivatives at the starting and finishing ends of the consecutive functions are equal.

The number of knots in the spline determines how many functions will compose the series. The degrees of the polynomial will determine the shape of the composing functions. The more functions, the more degrees of freedom, the more "waves". It is to the users' discretion to decide how smooth the function is supposed to be. A typical number of knots per year in a series of daily data would be in the range of 3–7 [12,39,40]. Katsouyanni et al. [41] suggest to choose the optimal number of knots by minimizing the absolute value of the partial autocorrelation of the residuals.

One must be in fact very careful when tailoring their spline function, as an excessive number of degrees of freedom or knots will lead to overfitting. An overfitted model is axiomatically not generalizable. An excessive number of degrees of freedom will make the model more "accurate", but also too specific to the training data sets. This might result in a model that is not applicable to other parametrically similar scenarios.

Combinations of Fourier terms and spline functions are not uncommon. Armstrong [42] used 6 harmonics to control

for seasonal patterns, and a natural cubic spline to capture the slower changes in trend throughout the time series.

Dummy variables for days

This is particularly important for time series aggregated in daily counts. It is not uncommon to control for days of the week when counting events like deaths, as it has been proven repeatedly that Mondays have a consistently higher count of deaths than other days. This can be done by incorporating dummy variables in the model indicating day of the week and holidays [43].

Other considerations that can be made is school days versus holidays, which can be connected to behaviors, infection rates etc. and hence should be controlled in some research questions. Such differentiations can be accounted for using Boolean variables, as exemplified by Todkill et al. [33]. An alternative that allows one to account for holidays is a form of seasonal decomposition used in economics is the "X11 decomposition", which follows the principle of classical decomposition, as it isolates the time series from its trend, seasonal and noise component, but it does so accounting for trading day variation, holidays and slow variations over the time series [44].

Lags

Accounting for lags is a fundamental step of a well-executed time series analysis. The effects of an extreme exposure do not manifest solely immediately respectively simultaneously with the event of exposure. For example, studies on the effects of extreme temperature on mortality consistently report that extremely low temperatures appear to have an effect on population mortality delayed over a number of days, whereas the effect of extremely high temperatures appears to be more acute [1].

Today's effect of an exposure that occurred in a previous day can be simply assessed by shifting the effector's measurement ahead on the series, that is, regressing the measured value on day t against the effect measured on day (t-lag number). This results in the creation of time-shifted copies of effect-exposure couples, which illustrate the effect of an independent variable at time t, over the response variable at time t+1,2,3.

This method however assumes that the series analyzed is stationary, which is often not the case. A stationary time series is defined as one whose points have the tendency to go back to its "long run" average [45]. To account for this, the weights of the lags can be distributed in such a way that, past a certain lag, the influence of a lagged value decreases, in a method called "restricted finite distributed lag model" (DLM). The principle behind this model is that the weights of the lagged values follow a predetermined shape, be it linear or quadratic. A more elegant and realistic approach involves accounting for the cumulative effect of exposure. Instead of accounting for the lags one at a time, lags could be accounted simultaneously, so that the lag effects are not confounded by each other. This can be obtained using DLMs. This methodology allows to determine the extent to which past variables of x influence y, by defying a clear pattern, whereas a simpler lag analysis illustrates the effects that a change in a variable at time t has on the response variable at time t+1.

An interesting methodology implemented by De Troeyer et al. [46], was to average out the independent variable measurements for a specific lag series, to get an idea about the overall exposure over a time period before the event occurred. As a downside, this approach gives very little information about the exposure and temporal threshold that triggered the event, and runs at a risk of hiding the accurate time fluctuations that led to the event.

One is advised to choose the most appropriate lag structure using empirical terms (statistical significance, Akaike information criteria, residual lags) [45]. Although this could result in a statistically sound model, the result might be lacking in mundane reliability. It has been commented that a seemingly biologically-arbitrary lag structure might be of very little use for informing environmental epidemi-

ologists. One must remember that at the end of the day, the real question that one tries to answer is an approximation of the extent that an exogenous variable affects a predetermined measure of population health, and not the mere solution to modelling and/or optimization exercise. Often, a third degree polynomial structure of the lag effects is preferred: Such a polynomial offers a maximum and a minimum (the first derivative has 2 zeros) that are easily interpreted as the lag with the strongest effect followed by a period dominated by the "harvesting effect". Of course, the situation gets even more complicated when the examined effect (at each lag) does not follow a linear dose-response shape. But also for DLNMs modern statistics programs now offer easy-to-use procedures.

As mentioned before, DLMs pose a solution to the problem of integrating lagged variables to the model. They are implemented to quantify the health effects of delayed exposure, accounting for harvesting [47]. In parallel to GAM and GLM, the main difference between DLM and DLNM is that DLNM is not limited to linear relationships, but rather bases itself on the cross matrix of linear and non-linear relationships between the different exposures (lagged and immediate) and the response.

Such an approach has been used by Todkill et al. [33] and Nayebare et al. [32]. Distributed lag non-linear models are easily implementable in R [48]. But as with any non-linear dose-response function, interpretation is less straightforward. While a linear approximation of a given relationship might not exactly represent the true situation, it can easily be summarized in a single risk estimate or coefficient.

Covariates

Seasonality and trend components are not the only sources of confounding noise in a time series. Time series analysis allows to compare different types and levels of exposure that variate with time within the same population, allowing for a natural control of ecosystem related confounding variables. It also renders the issue of individual based risk factors trivial (smoking, sex, age, etc.) as their distribution is assumed to remain stable over the course of the study. This level of control over external variables, together with the increased sample size, facilitates the detection of smaller relative risks that might be associated with exposure to pollutants [49].

There is to consider, however, that the ecosystem in which the sample population exists is not constant. As aforementioned, temperature and relative humidity are environmental variables that have been proved to have an effect on standard epidemiological measures (deaths, incidence of disease, etc.). The way to control for these extraneous variables is to include them into the model design.

Popular co-variates to control for when examining the impact of an air pollutant, are temperature, relative humidity, and other pollutants. As with the pollutant of interest, also these other factors might execute their effect over different lags and not necessarily following a linear dose-response relationship. For example, a steep change in temperature might be more stressful than a constantly high or low temperature. On the one hand, the researcher will not want to miss an important confounding effect. On the other hand, she might not spend too much study power and energy in the modelling of effects that are not central to the study question. Usually, an information criterion like the Akaike information criterion (AIC) [50] or the Bayesian information criterion (BIC) [51] will be used to select the optimal choice of co-variates and their lags.

In the end, it would be wise to first construct a model that controls:

– for long-term and seasonal variation,
– for meteorological factors.

The residuals of this model should mostly be normally distributed around zero with no sign of temporal trend or autocorrelation left. Only after these initial steps should the pollutant of interest tested against the residuals. Thus, the data would indicate when all confounders are optimally controlled for.

Linking the relevant time series

Regression analysis and model construction of count data can be implemented using GLM [52,53] or GAM, where Poisson is used as a link function. A link function is nothing but the paradigm against which the response variable is linked to its covariate.

The GLMs are a relatively simple way in which a linear model becomes a linear predictor of data.

The more primitive linear models describe the relationship between response variable and a predictor variable, assuming that the measurements of the response variable are independent, following a normal distribution. The GLM, on the other hand, ditches the assumption that the responses are coming from a normal distribution, and instead assumes that they are coming from any exponential family. More importantly, unlike linear regression, in GLM the slope between response and predictor variables can only be estimated via maximum likelihood. Rather than manipulation of the data, the idea is to manipulate the linear model so that the data can be analyzed as is, assuming an appropriate distribution which may not necessarily be normal.

More recent examples of this method being used are exemplified by Nayebare et al. [32] for an investigation of fine particle exposure and cardiopulmonary morbidity in Jeddah. Generalized linear models only accounts for linear components, which is why GAMs have been developed.

Generalized linear models are used to estimate smooth functional relationships between predictor variables and the response [54]. They are, if one wills, an organic evolution of GLM, in which the underlying functions describing the behavior of the datasets are not restricted to the linear model, and rather use flexible functions of time to estimate and account for overall trends and possible confounders in a non-parametric fashion. This allows the modeler to account for confounders that are non-linearly related to the response variable. It used the aforementioned splines to control for exposure covariates and trend, allowing one to build a detailed and flexible model.

Given the flexibility of this modelling method, it is often criticized for producing sample specific and non-generalizable models, which might represent a downside. Both GAMs and GLMs modelling methods can be implemented with ease in STATA, Python and R.

A less common method is the generalized estimating equation (GEE). This regression modelling method can be viewed as an extension of GLM. It offers a semi-parametric approach, as it does not fully specify the distribution of the generated data. Generalized estimating equations are used as a method for analyzing longitudinal, panel-type data [55], and assume, unlike the models mentioned above, that the responses of the model are correlated and non-independent from each other.

Generalized estimating equations estimate population average models, in which changes in the population mean are estimated based on the fluctuations of the covariates, and accounts for possible autocorrelation confounders [56].

This method assumes that the individuals within the same group will be more correlated than generalizable targets outside the sampled population (that is, the health outcome of a population within the panel will have more things in common than with the parameters of other populations). It has been used in the context of heat-related mortality by De Troeyer et al. [46].

A popular time series analysis method which arose in economics is the ARIMA model, which allows to model a time series and forecast it, whilst accounting for its trend by specifying for autoregressive and moving average terms. Its inherent weakness is that it does not include a seasonal component into the model, hence is not the model of choice for data where seasonal influences might play an important role. In ARIMA models the condition of stationarity is achieved by a series of differentiations [57].

The time series analysis can be made using the Box Jenkins method [58], which consist in 3 modelling stages: The first step is model identification by deciding which ARIMA model should be used taking account for the autocorrelation and partial autocorrelation scores. As a second step, the parameters that best fit the ARIMA are estimated algorithmically, and the third step is checking for residual independence, and constancy of residuals in variance and mean over time. The resulting model is used to forecast the testing data, and the fit was evaluated using the root mean squared errors.

Attempting to model any problem connected to a field is inherently a reductionist exercise of breaking down an issue into its manageable, quantifiable components. Typically, in environmental epidemiology, researchers are not so much interested in modelling seasonal and long-term trends and forecasting, but rather in finding associations indicative of causality. Therefore, ARIMA and similar models are usually not used in this field.

But still, in every scientific field we can learn from examining the models and techniques used in other fields.

The ways that have originated to answer the question of time reflect statistical methods are many and purpose-dependent. This is a rather obvious thought, but it is an important one to highlight when considering the translatability of a method from one discipline to another, or when choosing which method is the most appropriate to conduct one's analysis.

CONCLUSIONS

Time series analysis is a valuable tool for environmental epidemiology. Short-term effects of short-term exposures that are the main domain of TSA are often less important than chronic effects of prolonged exposures. Therefore, for example, impact assessments of air pollution usually are based on studies and data on chronic exposure [59]. Studies of the effects of chronic air pollution exposure usually depend on spatial contrasts that are often subjected to complex possibly confounding influences. Such studies therefore are usually much more cumbersome. And although there are some regions in the world with really low air pollution levels, populated areas experience on average at least some air pollution. Therefore, in spatial contrast, it is not possible to compare low pollution situations to an even lower, practically zero pollution. But with day-to-day variation in many settings days with almost zero pollution will occur. This does indeed enable the study of the impacts of low-level exposures as well. On the other hand, also extreme events (for example regarding meteorological conditions) are of particular interest. Impacts of extreme events are rather not studied by examining long-term averages.

Still, beware of over-interpretation of TSA results! Any TSA is just as good as its model parameters which need to be chosen well depending on the underlying research question. Often there is not a single true or correct approach. Every approach has its strengths and weaknesses. The following questions might sometimes help with the choice of the model:

- Are there probable seasonal influences that have to be adjusted? If so, is a simple adjustment enough, or should it be more flexible, e.g., using a Fourier series or a spline?
- Are there probable short-term influences that should be avoided using seasonal smoothing?
- Which covariates might probably represent confounding factors? How many should be included in the model?
- Does the data require a linear model, or is a non-linear model more plausible from a biological point of view?

But always keep in mind that predicting the future based on past trends is a vice of the naturalist. There are simply far too many variables to control, and reducing something as complex as population health as a number of covariables is reductionist, although useful.

And do not forget: the results of a TSA usually do not inform about the risk of individual persons but rather the "risk" of days (or other units of time) to experience a certain number

of "events". Take daily deaths as a typical example: to estimate the average risk of a person dying on a certain day you need to divide the number of deaths by the number of persons "at risk". But the latter number is only very crudely defined in TSA or often not even taken into account. Indeed, while in theory every person in a given area is at risk of death on every day, in truth the deaths mostly affect only a minority of all people in that area. Those prone to die are usually those most vulnerable, old and sick. And their number is usually not well defined. Thus, TSA informs about the relative risk in a poorly defined but highly relevant vulnerable subgroup only. This should be kept in mind especially when TSA results are used for impact assessments.

Author contributions

Research concept: Hanns Moshammer

Research methodology: Hanns Moshammer

Collecting material: Sandra Gudziunaite, Zana Shabani

Interpretation of results: Lisbeth Weitensfelder

References: Sandra Gudziunaite

REFERENCES

1. Weitensfelder L, Moshammer H. Evidence of Adaptation to Increasing Temperatures. Int J Environ Res Public Health. 2020;17:97. https://doi.org/10.3390/ijerph17010097
2. Kioumourtzoglou MA, Raz R, Wilson A, Fluss R, Nirel R, Broday DM, Yuval, et al. Traffic-related air pollution and pregnancy loss. Epidemiology. 2019;30:4-10. https://doi.org/10.1097/EDE.0000000000000918
3. Gudziunaite S, Moshammer H. Temporal patterns of weekly births and conceptions predicted by meteorology, seasonal variation, and lunar phases. Wiener Klinische Wochenschrift. 2022;134(13-14):538-545. https://doi.org/10.1007/s00508-022-02038-7
4. Page LA, Hajat S, Kovats RS. Relationship between daily suicide counts and temperature in England and Wales. Br J Psychiatry. 2007;191:106-112. https://doi.org/10.1192/bjp.bp.106.031948
5. Lehmann F, Alary PE, Rey G, Slama R. Association of Daily Temperature with Suicide Mortality: A Comparison With Other Causes of Death and Characterization of Possible Attenuation Across 5 Decades. Am J Epidemiol. 2022;191(12): 2037-2050. https://doi.org/10.1093/aje/kwac150
6. Pincus SM, Goldberger AL. Physiological time-series analysis: what does regularity quantify? Am J Physiol Heart Circ Physiol. 1994;266:H1643–H1656. https://doi.org/10.1152/ajpheart.1994.266.4.H1643
7. Franses PH, Koehler AB. A model selection strategy for time series with increasing seasonal variation. Int J Forecast 1998;14(3):405-414. https://doi.org/10.1016/S0169-2070(98)00041-7
8. Cheng BS. Energy consumption and economic growth in Brazil, Mexico and Venezuela: a time series analysis. Appl Econ Lett 1997;4:671-674. https://doi.org/10.1080/758530646.
9. Trancart T, Acou A, Oliveira ED, Feunteun E. Forecasting animal migration using SARIMAX: an efficient means of reducing silver eel mortality caused by turbines. Endanger Spec Res 2013;21:181-190. https://doi.org/10.3354/esr00517.
10. Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet. 2012;13(8):552-564. https://doi.org/10.1038/nrg3244
11. Parzen E. ARARMA models for time series analysis and forecasting. J Forecast 1982;1(1);67-82. https://doi.org/10.1002/for.3980010108
12. Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B. Time series regression studies in environmental epidemiology. Int J Epidemiol. 2013;42(4):1187-1195. https://doi.org/10.1093/ije/dyt092
13. Wen Q, Gao J, Song X, et al. RobustSTL: A Robust Seasonal-Trend Decomposition Algorithm for Long Time Series. Proceedings of the AAAI Conference on Artificial Intelligence 2019;33:5409-5416. https://doi.org/10.1609/aaai.v33i01.33015409

14. Schwartz J, Marcus A. Mortality and Air Pollution J London: A Time Series Analysis. Am J Epidemiol. 1990;131:185–194. https://doi.org/10.1093/oxfordjournals.aje.a115473.
15. Neuberger M, Rabczenko D, Moshammer H. Extended effects of air pollution on cardiopulmonary mortality in Vienna. Atmos Environ. 2007 41: 8549–8556. https://doi.org/10.1016/j.atmosenv.2007.07.013
16. Gasparrini A, Armstrong B. The impact of heat waves on mortality. Epidemiology. 2011;22(1):68–73. https://doi.org/10.1097/EDE.0b013e3181fdcd99
17. Heo S, Lee W, Bell ML. Suicide and Associations with Air Pollution and Ambient Temperature: A Systematic Review and Meta-Analysis. Int J Environ Res Public Health. 2021;18(14):7699. https://doi.org/10.3390/ijerph18147699.
18. Shah AV, Lee KK, McAllister DA, Hunter A, Nair H, Whiteley W, et al. Short term exposure to air pollution and stroke: systematic review and meta-analysis. BMJ. 2015;350:h1295. https://doi.org/10.1136/bmj.h1295
19. Wang Y, Eliot MN, Wellenius GA. Short-term changes in ambient particulate matter and risk of stroke: A systematic review and meta-analysis. J Am Heart Assoc. 2014; 3:e000983. https://doi.org/10.1161/JAHA.114.000983
20. Gu J, Shi Y, Chen N, Wang H, Chen T. Ambient fine particulate matter and hospital admissions for ischemic and hemorrhagic strokes and transient ischemic attack in 248 Chinese cities. Sci Total Environ. 2020; 715:136896. https://doi.org/10.1016/j.scitotenv.2020.136896.
21. Chen C, Liu X, Wang X, Qu W, Li W, Dong L. Effect of air pollution on hospitalization for acute exacerbation of chronic obstructive pulmonary disease, stroke, and myocardial infarction. Environ Sci Pollut Res. 2020;27:3384–3400. https://doi.org/10.1007/s11356-019-07236-x.
22. Guo Y, Xie X, Lei L, Zhou H, Deng S, Xu Y, et al. Short-term associations between ambient air pollution and stroke hospitalisations: time-series study in Shenzhen, China. BMJ Open. 2020;10:32974. https://doi.org/10.1136/bmjopen-2019-032974
23. Slama A, Śliwczyński A, Woźnica-Pyzikiewicz J, Zdrolik M, Wisnicki B, Kubajek J, et al. The short-term effects of air pollution on respiratory disease hospitalizations in 5 cities in Poland: comparison of time-series and case-crossover analyses. Environ Sci Pollut Res. 2020;27:24582–24590. https://doi.org/10.1007/s11356-020-08542-5
24. Ashworth M, Analitis A, Whitney D, Samoli E, Zafeiratou S, Atkinson R, et al. Spatio-temporal associations of air pollutant concentrations, GP respiratory consultations and respiratory inhaler prescriptions: a 5-year study of primary care in the borough of Lambeth, South London. Environ Health. 2021;20(1):54. https://doi.org/10.1186/s12940-021-00730-1
25. Shabani Isenaj Z, Berisha M, Ukëhaxhaj A, Moshammer H. Particulate Air Pollution and Primary Care Visits in Kosovo: A Time-Series Approach. Int J Environ Res Public Health. 2022;19(24):16591. https://doi.org/10.3390/ijerph192416591
26. Imai C, Armstrong B, Chalabi Z, Mangtani P, Hashizume M. Time series regression model for infectious disease and weather. Environ Res. 2015;142:319-327. https://doi.org/10.1016/j.envres.2015.06.040
27. Bercu S, Proïa F. A SARIMAX coupled modelling applied to individual load curves intraday forecasting. J Appl Stat. 2013;40:1333–1348. https://doi.org/10.1080/02664763.2013.785496
28. Schwartz J, Spix C, Touloumi G, Bachárová L, Barumamdzadeh T, le Tertre A, et al. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. J Epidemiol Community Health. 1996;50(Suppl 1): S3-S11. https://doi.org/10.1136/jech.50.suppl_1.s3
29. Saputro DRS, Susanti A, Pratiwi NBI. The handling of overdispersion on Poisson regression model with the generalized Poisson regression model. AIP Conference Proceedings. 2021;2326:020026. https://doi.org/10.1063/5.0040330
30. Onozuka D, Hagihara A. Variation in vulnerability to extreme-temperature-related mortality in Japan: A 40-year time-series analysis. Environ Res. 2015;140:177-184. https://doi.org/10.1016/j.envres.2015.03.031
31. Gu J, Shi Y, Zhu Y, Chen N, Wang H, Zhang Z, et al. Ambient air pollution and cause-specific risk of hospital

admission in China: A nationwide time-series study. PLoS Med. 2020;17(8):e1003188. https://doi.org/10.1371/journal.pmed.1003188

32. Nayebare SR, Aburizaiza OS, Siddique A, Carpenter DO, Pope CA, Mirza HM, et al. Fine particles exposure and cardiopulmonary morbidity in Jeddah: A time-series analysis. Sci Total Environ. 2019;647:1314-1322. https://doi.org/10.1016/j.scitotenv.2018.08.094
33. Todkill D, de Jesus Colon Gonzalez F, Morbey R, Charlett A, Hajat S, Kovats S, et al. Environmental factors associated with general practitioner consultations for allergic rhinitis in London, England: a retrospective time series analysis. BMJ Open. 2020;10(12):e036724. https://doi.org/10.1136/bmjopen-2019-036724
34. Moshammer H, Poteser M, Kundi M, Kemmerer K, Weitensfelder L, Wallner P, et al. Nitrogen-Dioxide Remains a Valid Air Quality Indicator. Int J Environ Res Public Health. 2020;17(10):3733. https://doi.org/10.3390/ijerph17103733
35. Poteser M, Moshammer H. Daylight Saving Time Transitions: Impact on Total Mortality. Int J Environ Res Public Health. 2020;17(5):1611. https://doi.org/10.3390/ijerph17051611
36. Aik J, Ong J, Ng LC. The effects of climate variability and seasonal influence on diarrhoeal disease in the tropical city-state of Singapore – A time-series analysis. Int J Hyg Environ Health. 2020;227:113517. https://doi.org/10.1016/j.ijheh.2020.113517
37. Cleveland RB, Cleveland WS, McRae JE, Terpenning I. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. J Off Stat. 1990;6(1):3-73.
38. Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. BMC Med Res Methodol. 2019;19:46. https://doi.org/10.1186/s12874-019-0666-3
39. Dominici F, McDermott A, Zeger SL, Samet JM. On the use of generalized additive models in time-series studies of air pollution and health. Am J Epidemiol. 2002;156(3):193-203. https://doi.org/10.1093/aje/kwf062
40. Dominici F, Samet J M, Zeger SL. Combining Evidence on Air pollution and Daily Mortality from the Twenty Largest US cities: A Hierarchical Modelling Strategy. Royal Statistical Society, Series A: Statistics in Society 2000;163:263-302. https://doi.org/10.1111/1467-985X.00170
41. Katsouyanni K, Schwartz J, Spix C, Touloumi G, Zmirou D, Zanobetti A, et al. Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. J Epidemiol Community Health. 1996;50 Suppl 1(Suppl 1):S12-S18. https://doi.org/10.1136/jech.50.suppl_1.s12
42. Armstrong B. Models for the relationship between ambient temperature and daily mortality. Epidemiology. 2006;17(6):624-631. https://doi.org/10.1097/01.ede.0000239732.50999.8f
43. Tapia V, Steenland K, Sarnat SE, Vu B, Liu Y, Sánchez-Ccoylloet O, al. Time-series analysis of ambient PM2.5 and cardiorespiratory emergency room visits in Lima, Peru during 2010-2016. J Expo Sci Environ Epidemiol. 2020;30(4):680-688. https://doi.org/10.1038/s41370-019-0189-3
44. Dagum EB, Bianconcini S. Seasonal adjustment methods and real time trend-cycle estimation. Springer International, Switzerland; 2016. https://doi.org/10.1007/978-3-319-31822-6
45. Shrestha M, Bhatta G. Selecting appropriate methodological framework for time series data analysis. J Financ Data Sci. 2018; 4:71-89. https://doi.org/10.1016/j.jfds.2017.11.001
46. De Troeyer K, Bauwelinck M, Aerts R, Profer D, Berckmans J, Delcloo A, et al. Heat related mortality in the two largest Belgian urban areas: A time series analysis. Environ Res. 2020;188:109848. https://doi.org/10.1016/j.envres.2020.109848
47. Gasparrini A, Armstrong B, Kenward MG. Distributed lag non-linear models. Statist Med. 2010;29:2224-2234. https://doi.org/10.1002/sim.3940
48. Gasparrini A, Scheipl F, Armstrong B, Kenward MG. A penalized framework for distributed lag non-linear models [published correction appears in Biometrics. 2022;78(2):812]. Biometrics. 2017;73(3):938-948. https://doi.org/10.1111/biom.12645

49. Pekkanen J, Pearce N. Environmental epidemiology: challenges and opportunities. Environ Health Perspect. 2001;109(1):1-5. https://doi.org/10.1289/ehp.011091
50. Akaike, H. Information theory as an extension of the maximum likelihood principle. In: Petrov BN, Csaksi F, editors. Proceedings of the Second International Symposium on Information Theory, Tsahkadsor, Armenia, 2–8 September 1971; Akademiai Kiado: Budapest, Hungary.
51. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978; 6:461-4. https://doi.org/10.1214/aos/1176344136
52. Chuang Y, Mazumdar S, Park T, Tang G, Arena VC, Nicolich MJ. Generalized linear mixed models in time series studies of air pollution. Atmos Pollut Res. 2011;2(4):428-435. https://doi.org/10.5094/APR.2011.049
53. Crosbie SF, Hinch GN. An intuitive explanation of generalised linear models. New Zealand J Agric Res. 1985;28: 19-29. https://doi.org/10.1080/00288233.1985.10426995
54. Pedersen EJ, Miller DL, Simpson GL, Ross N. Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ. 2019;7:e6876. https://doi.org/10.7717/peerj.6876
55. Saez M, Tobias A, Muñoz P, Campbell MJ. A GEE moving average analysis of the relationship between air pollution and mortality for asthma in Barcelona, Spain. Stat Med. 1999;18(16):2077-2086. https://doi.org/10.1002/(sici)1097-0258(19990830)18:16%3c2077::aid-sim185%3e3.0.co;2-t.
56. Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, Jewell N, et al. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology. 2010;21(4):467-474. https://doi.org/10.1097/EDE.0b013e3181caeb90
57. Miyama T, Matsui H, Azuma K, Minejima C, Itano Y, Takenaka N, et al. Time Series Analysis of Climate and Air Pollution Factors Associated with Atmospheric Nitrogen Dioxide Concentration in Japan. Int J Environ Res Public Health. 2020;17(24):9507. https://doi.org/10.3390/ijerph17249507
58. Box G, Jenkins G. Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.
59. WHO. Health Risks of Air Pollution in Europe – HRAPIE Project. 2017. Accessed on 2022 December 28, available from: https://www.who.int/europe/publications/i/item/WHO-EURO-2013-6696-46462-67326