‘eesim’: R Software for Environmental Epidemiology Simulations

Sarah Koehler and Brooke Anderson, Colorado State University

Introduction

The ‘eesim’ package provides functions to create simulated time series of environmental exposures (e.g., temperature, air pollution) and health outcomes for use in power analysis and simulation studies in environmental epidemiology. This package also provides functions to evaluate the results of simulation studies based on these simulated time series.

Motivation

Simulation studies are important in environmental epidemiology research on air pollution, temperature, and other exposures. For example, simulated data can be used to test new statistical models and perform power analyses. Simulated data has been used in a number of studies in environmental epidemiology:

- Investigating short-term mortality displacement following heat waves (Armstrong 2014)
- Comparing experimental designs and model choice (Bateson 1999; Bateson 2001; Peng 2006; Roberts 2006)
- Investigating geographic heterogeneity of exposures (Strickland 2015; Gryparis 2009; Butland 2013)
- Assessing the performance of a proposed method for estimating the health effects of multi-pollutant exposures (Bobb 2015).

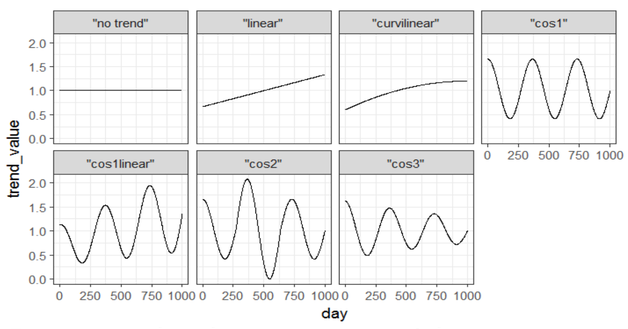
Two current challenges of simulating time series data for epidemiology studies are:

(1) Methods for simulating are inconsistent across studies
(2) It’s time-consuming to develop code to simulate environmental health time series.

Customization

An important feature of `eesim` is that the user can create and use custom functions for any part of the simulation process. Functions the user has the option to customize within the `eesim` framework are:

- Exposure trend
- Outcome trend
- How exposure influences the expected outcomes
- Randomization from the trend lines
- Models for fitting the simulated data



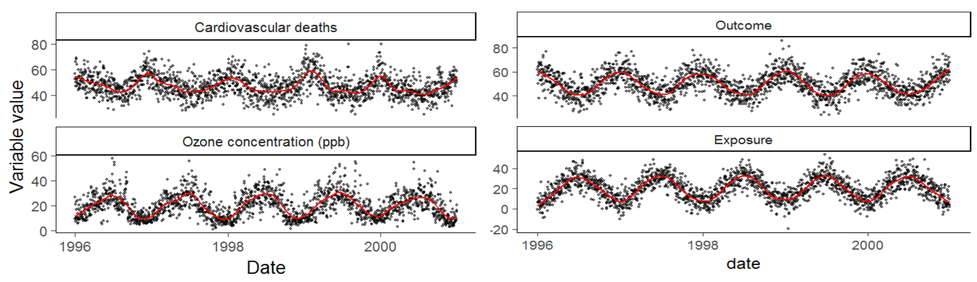
For example, the user may wish to generate exposure data with a custom trend, then automate the processes of generating outcomes, fitting models, and evaluating performance using the built-in features of `eesim`, such as the trends shown above.

Assessing model performance

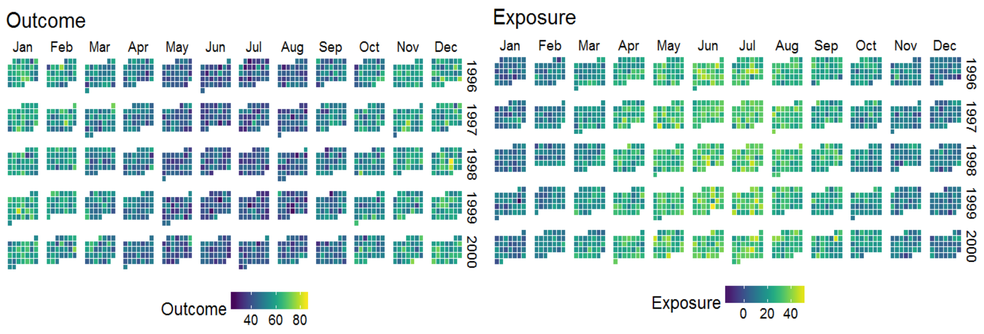
Here is an example of using `eesim` to assess model performance for an analysis of associations between ozone and mortality. The `eesim` package facilitates four steps in this process:

1. Generation of exposure data;
2. Generation of outcome data;
3. Fitting models to simulated data; and
4. Evaluating model performance on simulated data.

The plot on the left shows time series of daily ozone concentration (in parts per billion [ppb]) and cardiovascular deaths in Chicago, IL (1996–2000). We used the characteristics of this data to simulate similar datasets (example on right) to evaluate model performance.



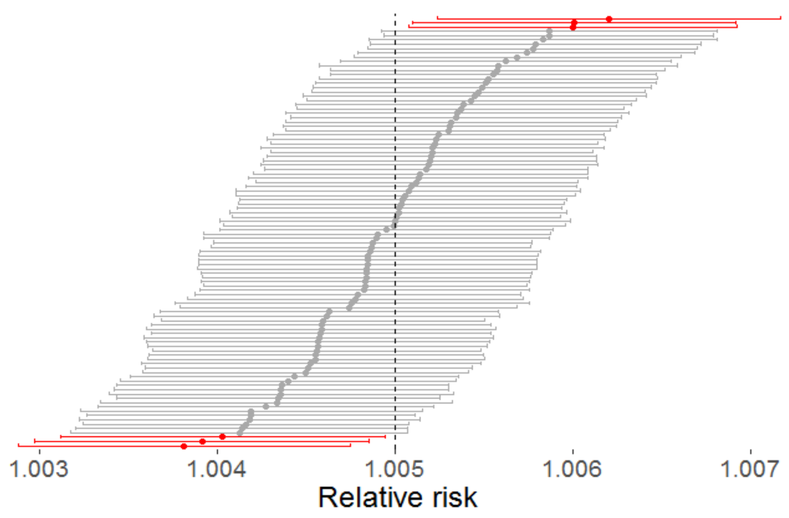
The package also allows visualization of this data through the `calendar_plot`:



The `eesim` function generates multiple similar simulated datasets to investigate model performance, returning several summaries of model performance across all simulations:

Variable	Description
`beta_hat`	**Mean estimate**: The mean of the estimated log relative rate over all simulations.
`rr_hat`	**Mean estimated relative rate**: The mean of the estimated relative rate over all simulations.
`var_across_betas`	**Variance across estimates**: Variance of the point estimates (estimated log relative risk) over all simulations.
`mean_beta_var`	**Mean variance of estimate**: The mean of the variances of the estimated effect (estimated log relative risk) across all simulations.
`percent_bias`	**Relative bias**: Difference between the estimated log relative risk and true log relative risk as a proportion of the true log relative risk.
`coverage`	**95% confidence inverval coverage**: Percent of simulations for which the 95% confidence interval estimate of log relative risk includes the true value of log relative risk.
`power`	**Power**: Percent of simulations for which the null hypothesis that the log relative risk equals zero is rejected based on a p-value of 0.05.

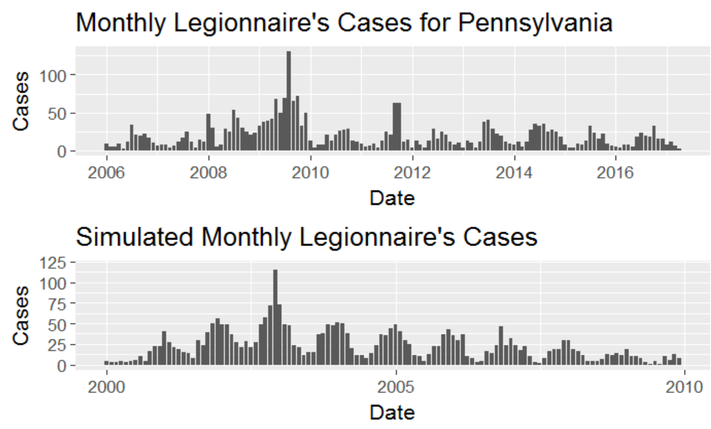
After running the simulation, you can look at the relative risk point estimate and 95% confidence interval from each of the 100 simulations, as well as which 95% confidence intervals include the true relative rate, using the `coverage_plot` function that comes with the package.



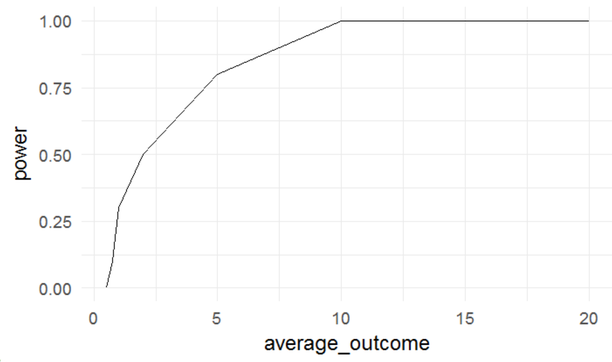
Power analysis

Simulation studies can be used to estimate effective sample size or power when planning or proposing future research studies, especially when a particularly complex model will be needed for analysis and when the strong assumptions of classical, analytical power analysis are questionable (Bellan 2015; Johnson 2015).

Here is an example of conducting a power analysis to study the relationship between Legionnaires' disease (LD) and extreme precipitation in a community (Brandsema 2014).



The `eesim` package can use the simulated data to estimate power for different sample sizes, relative risks, or average outcomes. Since Legionnaires’ disease is relatively rare, we might be most interested in the power resulting from different average daily case counts, which would help determine if enough cases occur in a community to allow an adequately-powered analysis. The plot below is returned from the `power_calc` function in `eesim`.



References

Armstrong B, Gasparrini A, Hajat S (2014). “Estimating mortality displacement during and after heat waves.” *American Journal of Epidemiology*, **179**(12), 1405-1406.doi: 10.1093/aje/kwu083.

Bateson TF, Schwartz J (1999). “Control for seasonal variation and time trend in case-crossover studies of acute effects of environmental exposures.” *Epidemiology*, **10**(5), 539-544.

Bateson TF, Schwartz J (2001). “Selection bias and confounding in case-crossover analyses of environmental time-series data.“ *Epidemiology*, **12**(6), 654-661.

Bellan SE, Pulliam JR, Pearson CA, Champredon D, Fox SJ, Skrip L, Galvani AP, Gambhir M, Lopman BA, Porco TC, et al. (2015). “Statistical power and validity of Ebola vaccine trials in Sierra Leone: a simulation study of trial design and analysis.” *The Lancet Infectious Diseases*, **15**(6), 703-710.

Bobb JF, Valeri L, Henn BC, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA (2015). “Bayesian kernel machine regression for estimating the health effects of multipollutant mixtures.” *Biostatistics*, **16**(3), 493-508. doi:10.1093/biostatistics/kxu058.

Brandsema, P. S., Euser, S. M., Karagiannis, I., Den Boer, J. W., & Van Der Hoek, W. (2014). “Summer increase of Legionnaires' disease 2010 in The Netherlands associated with weather conditions and implications for source finding.” *Epidemiology and infection*, **142**(11), 2360-2371.

Butland BK, Armstrong B, Atkinson RW, Wilkinson P, Heal MR, Doherty RM, Vieno M (2013). “Measurement error in time-series analysis: a simulation study comparing modelled and monitored data.” *BMC Medical Research Methodology*, **13**, 136.

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA (2009). “Measurement error caused by spatial misalignment in environmental epidemiology.” *Biostatistics*, **10**(2), 258-274. doi:10.1093/biostatistics/kxn033.

Johnson PC, Barry SJ, Ferguson HM, Müller P (2015). “Power analysis for generalized linear mixed models in ecology and evolution.” *Methods in Ecology and Evolution*, **6**(2), 133-142.

Peng RD, Dominici F, Louis TA (2006). “Model choice in time series studies of air pollution and mortality.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**(2), 179-203.

Roberts S, Martin MA (2006). “The question of nonlinearity in the dose-response relation between particulate matter air pollution and mortality: Can Akaike's Information Criterion be trusted to take the right turn?” *American Journal of Epidemiology*, **164**(164), 1242-1250. doi:10.1093/aje/kwj335.

Strickland MJ, Gass KM, Goldman GT, Mulholland JA (2015). “Effects of ambient air pollution measurement error on health effect estimates in time-series studies: a simulation-based analysis.” *Journal of Exposure Science and Environmental Epidemiology*, **25**(2), 160-166.

Acknowledgements: *This work was supported by a grant from the National Institute of Environmental Health Sciences (R00ES022631) and a fellowship from the Colorado State University Programs for Research and Scholarly Excellence.*