The Journal of Open Source Software

APCtools: Descriptive and Model-based Age-Period-Cohort Analysis

Alexander Bauer[1], **Maximilian Weigert**[1], **and Hawre Jalal**[2]

1 Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Germany **2** Department of Health Policy and Management, Graduate School of Public Health, University of Pittsburgh.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Nikoleta Glynatsi

Reviewers:

- @strengejacke
- @jaybee84

Submitted: 02 January 2022
Published: unpublished

Summary

Age-Period-Cohort (APC) analysis aims to determine relevant drivers for long-term developments and is used in many fields of science (Yang & Land, 2013). The R package `APCtools` offers modern visualization techniques and general routines to facilitate the interpretability of the interdependent temporal structures and to simplify the workflow of an APC analysis. Separation of the temporal effects is performed utilizing a semiparametric regression approach. We shortly discuss the challenges of APC analysis, give an overview of existing statistical software packages and outline the main functionalities of the package.

Statement of Need

The main focus in APC analysis is on disentangling the interconnected effects of age, period, and cohort. Long-term developments of some characteristic can either be associated with changes in a person's life cycle (age), macro-level developments over the years that simultaneously affect all age groups (period), or the generational membership of an individual, shaped by similar socialization processes and historical experiences (cohort).

The critical challenge in APC analysis is to deal with the perfect linear dependency of the components age, period, and cohort (cohort = period - age). Due to this *identification problem*, inferring on the actual drivers behind observed temporal developments is difficult. For example, changes in recent years could be explained by developments related to the period, or by the fact that the respective observations only comprise later cohorts. The estimation of a linear regression model with all three components as individual main effects is only possible when imposing additional constraints in the estimation process, like restricting one main effect to zero (Yang & Land, 2013). Such explicit constraints, however, typically result in effect structures that are hard to interpret. Accordingly, flexible methods and visualization techniques are needed that rely on less restrictive assumptions to circumvent the identification problem.

Several packages for APC analysis exist for the statistical software R. Package apc (Fannon & Nielsen, 2020) implements methods based on the canonical parametrization of Kuang et al. (2008), which however lack flexibility and robustness when compared to nonlinear regression approaches. Package `bamp` (Schmid & Held, 2007) offers routines for the analysis of incidence and mortality data based on a Bayesian APC model with a nonlinear prior. R package `Epi` (Carstensen et al., 2021) implements the methods introduced in Carstensen (2007) to analyze disease and mortality rates, including the estimation of separate smooth effects for age, period and cohort. Rosenberg et al. (2014) developed an R-based web tool for the analysis of cancer rates, including different estimates for marginal effect curves.

In contrast to the above software packages, `APCtools` builds on a flexible and robust semiparametric regression approach. The package includes modern visualization techniques and general

The Journal of Open Source Software

routines to facilitate the interpretability of the estimated temporal structures and to simplify the workflow of an APC analysis. As is outlined below in further detail, sophisticated functions are available both for descriptive and regression model-based analyses. For the former, we use density (or ridgeline) matrices, classical heatmaps and *hexamaps* (hexagonally binned heatmaps) as innovative visualization techniques building on the concept of Lexis diagrams. Model-based analyses build on the separation of the temporal dimensions based on generalized additive models, where a tensor product interaction surface (usually between age and period) is utilized to represent the third dimension (usually cohort) on its diagonal. Such tensor product surfaces can also be estimated while accounting for further covariates in the regression model.

Descriptive Analysis

In the following, we showcase the main functionalities of the `APCtools` package on the included `travel` dataset, containing data from the German *Reiseanalyse* survey (Forschungsgemeinschaft Urlaub und Reisen e.V., 2022) – a repeated cross-sectional study comprising information on German travelers between 1971 and 2018. Focus is on travelers between 14 and 89 years and the distance of each traveler's *main trip* – i.e. each traveler's most important trip in the respective year – and how these distances change over the temporal dimensions.

Several descriptive visualization techniques are implemented that are all based on the classical concept of Lexis diagrams where two temporal dimensions (of age, period, and cohort) are depicted on the x- and y-axis, and the remaining dimension along the diagonals. Additional to heatmaps and *hexamaps* (see below) this includes density matrices (called *ridgeline matrices* in Weigert et al. (2021)) which can be used to flexibly visualize observed distributions along the temporal dimensions. Such visualizations can for example be used to illustrate changes in travel distances. As can be seen in Figure 1 and Figure 3, longer-distance travels are mainly undertaken by young age groups and in more recent years.

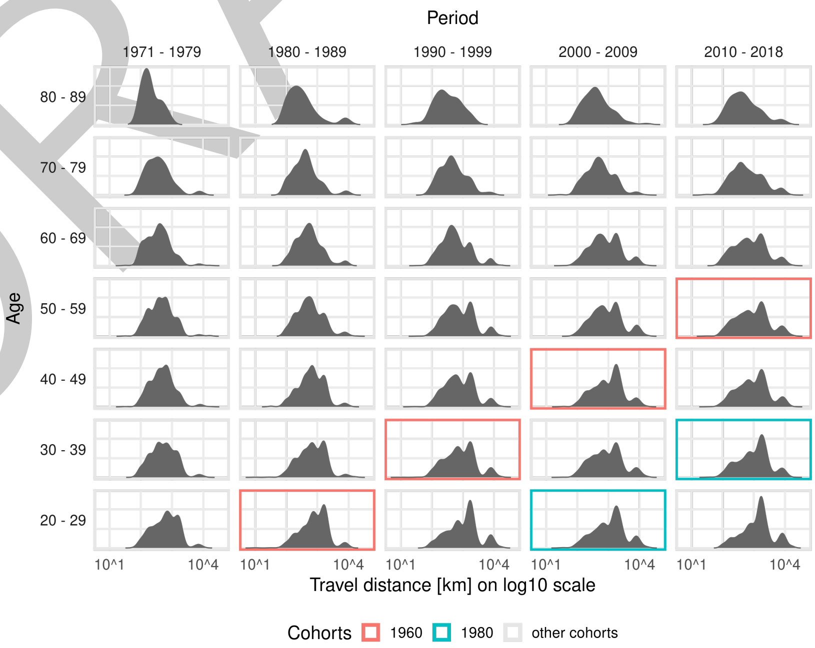

Figure 1: Density matrix of the main trips' travel distance in different age and period groups. Two cohort groups are exemplarily highlighted.

The Journal of Open Source Software

Model-based Analysis

To properly estimate the association of a process with the individual dimensions age, period, and cohort, we utilize the approach introduced by Clements et al. (2005) who circumvent the identification problem by representing the effect of one temporal dimension (e.g. cohort) based on a nonlinear interaction surface between the other two dimensions (age and period). This leads to a generalized additive regression model (GAM, Wood (2017)) of the following form:

$$g(\mu_i) = \beta_0 + f_{ap}(age_i, period_i) + \eta_i, \qquad i = 1, \ldots, n,$$

with observation index i, μ_i the expected value of an exponential family response, link function $g(\cdot)$ and the intercept β_0. The interaction surface is included as a tensor product surface $f_{ap}(age_i, period_i)$, represented by a two-dimensional spline basis. η_i represents an optional linear predictor that contains further covariates. Model estimation can be performed with functions `gam` or `bam` from R package `mgcv` (Wood, 2017). As outlined in Weigert et al. (2021) this modeling approach can both be applied to repeated cross-sectional data and panel data.

Based on an estimated GAM, a heatmap of the smooth tensor product surface can be plotted (see Figure 2). Additionally, marginal effects of the individual temporal dimensions can be extracted by averaging over each dimension.

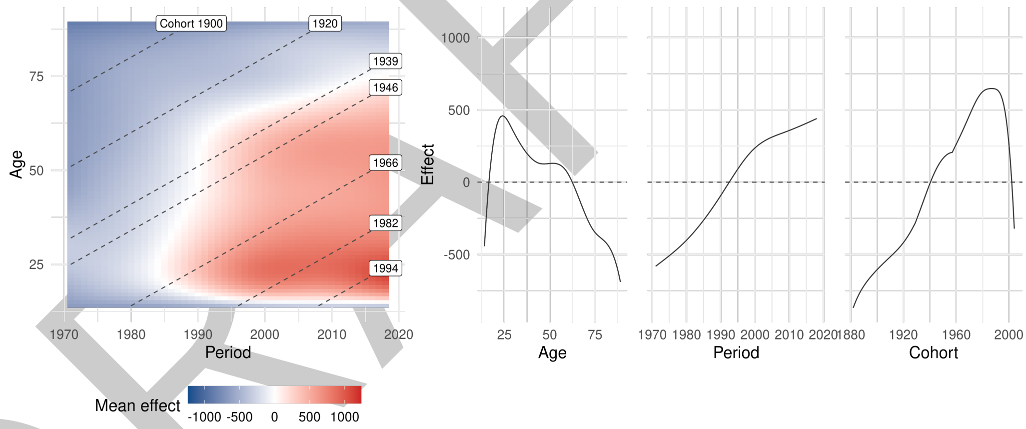

Figure 2: Heatmap of the estimated tensor product surface (left pane) and marginal APC effects based on an additive model with the travel distance as response and no further control variables (right pane).

As an alternative to classical heatmaps the raw observed APC structures or the subsequently estimated model-based tensor product surface can also be visualized using *hexamaps*, i.e. hexagonally binned heatmaps where developments over age, period, and cohort are given equal visual weight by distorting the coordinate system (Jalal & Burke, 2020). This resolves the central problem of classical heatmaps where developments over the diagonal dimension are visually underrepresented compared to developments over the dimensions depicted on the x- and y-axis.

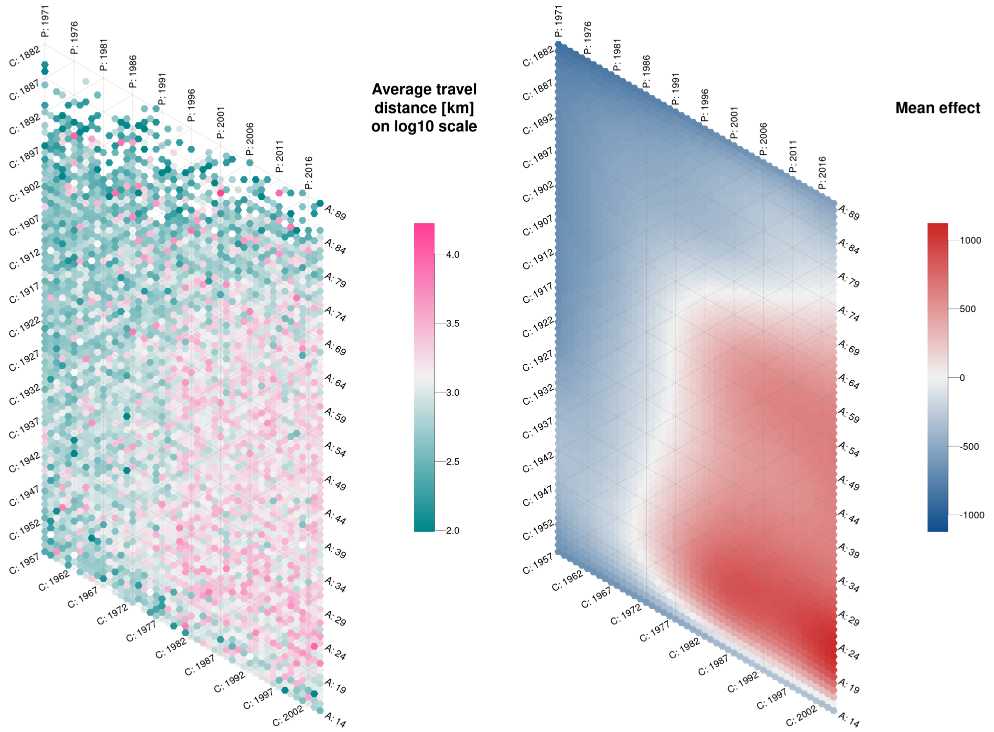

Figure 3: Hexamaps of the observed travel distances (left pane) and the estimated tensor product surface based on an additive model with the travel distance as response and no further control variables (right pane).

`APCtools` further provides partial APC plots, which can be used to visualize interdependencies between the different temporal dimensions (see Weigert et al. (2021) for details). Also, several utility functions are available to plot covariate effects as well as functions to create publication-ready summary tables of the central model results.

Acknowledgments

We thank Helmut Küchenhoff for valuable methodological contributions. This work has been partially funded by the German Research Foundation (DFG) under Grant No. KU 1359/4-1 and by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A.

References

Carstensen, B. (2007). Age–period–cohort models for the lexis diagram. *Statistics in Medicine*, *26*(15), 3018–3045. https://doi.org/10.1002/sim.2764

Carstensen, B., Plummer, M., Laara, E., & Hills, M. (2021). *Epi: A package for statistical analysis in epidemiology*. https://CRAN.R-project.org/package=Epi

Clements, M. S., Armstrong, B. K., & Moolgavkar, S. H. (2005). Lung cancer rate predictions using generalized additive models. *Biostatistics*, *6*(4), 576–589. https://doi.org/10.1093/biostatistics/kxi028

Fannon, Z., & Nielsen, B. (2020). *Apc: Age-period-cohort analysis*. https://CRAN.R-project.org/package=apc

Forschungsgemeinschaft Urlaub und Reisen e.V. (2022). *Survey of tourist demand in germany for holiday travel and short breaks*. https://reiseanalyse.de/wp-content/uploads/2021/06/Reiseanalyse-2022_Informationsbroschuere.pdf

Jalal, H., & Burke, D. S. (2020). Hexamaps for age-period-cohort data visualization and implementation in R. *Epidemiology (Cambridge, Mass.)*, *31*(6), e47. https://doi.org/10.1097/EDE.0000000000001236

Kuang, D., Nielsen, B., & Nielsen, J. P. (2008). Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, *95*(4), 979–986. https://doi.org/10.1093/biomet/asn026

Rosenberg, P. S., Check, D. P., & Anderson, W. F. (2014). A web tool for age–period–cohort analysis of cancer incidence and mortality rates. *Cancer Epidemiology, Biomarkers & Prevention*, *23*, 2296–2302. https://doi.org/10.1158/1055-9965.EPI-14-0300

Schmid, V. J., & Held, L. (2007). BAMP – bayesian age-period-cohort modeling and prediction. *Journal of Statistical Software*, *21*. https://doi.org/10.18637/jss.v021.i08

Weigert, M., Bauer, A., Gernert, J., Karl, M., Nalmpatian, A., Küchenhoff, H., & Schmude, J. (2021). Semiparametric APC analysis of destination choice patterns: Using generalized additive models to quantify the impact of age, period, and cohort on travel distances. *Tourism Economics*, 1354816620987198. https://doi.org/10.1177/1354816620987198

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC press. https://doi.org/10.1201/9781315370279

Yang, Y., & Land, K. C. (2013). *Age-period-cohort analysis: New models, methods, and empirical applications*. Taylor & Francis. https://doi.org/10.1201/b13902