**README for GitHub repository accompanying the paper:**

# Augmenting the availability of historical GDP per capita estimates through machine learning

**by Philipp Koch, Viktor Stojkoski, & César A. Hidalgo**

## Version updates

*v2: Now including source data for European regions in 1900 and 1950 provided by Rosés & Wolf.*[1,2] *We rescale their estimates such that they match country-level data by the Maddison project. This leads to slightly different out-of-sample estimates in 1900 and 1950 than in the original version. The repository, thus, also includes an Excel file with the original version 'historicalGDPs_augmented_dataset_v1.xlsx' and information in the code '00_master.R' on how to replicate v1.*

The purpose of this document is twofold. First, we want to provide explanations for the final dataset and results that we publish. Second, we want to provide detailed information on the code to ensure replicability.

For all questions regarding the repository, please contact Philipp Koch (philipp.koch@ecoaustria.ac.at).

## Dataset & results

The final dataset (*historicalGDPs_augmented_dataset.xlsx*) contain 5,700 observations with the following columns:

- *ID*: location_year
- *location*: denotes the code for the respective country (iso-3) or region. The classification of regions follows the NUTS-2 (2021 edition) classification in Europe, metro- and micropolitan areas in the US, metropolitan areas in Canada, and oblasts in Ukraine and Russia.
- *country_0*: This column is relevant, if the location is a region. Then country_0 denotes the country the region is in.
- *location_name*: denotes the name of the respective country or region.
- *year*: in 50-year intervals
- *GDPpc*: GDP per capita levels denoted in 2011 USD.

[1] Rosés, J. R., & Wolf, N. (2021). Regional growth and inequality in the long-run: Europe, 1900–2015. *Oxford Review of Economic Policy*, *37*(1), 17–48. https://doi.org/10.1093/oxrep/graa062

[2] Rosés, J. R., & Wolf, N. (2019). *The economic development of Europe's regions: A quantitative history since 1900*. Routledge.

- *flag*: denotes whether an observation is based on source data, or an out-of-sample estimate based on our machine learning models. Also, it denotes the specific source.
- *GDPpc_lower* & *GDPpc_upper*: 90% confidence interval for out-of-sample estimates obtained by bootstrapping.

The dataset contains 5,700 observations, with 1,574 observations from source data and 4,126 out-of-sample estimates. All references for source data observations are described in detail in the manuscript and the SI Appendix.

The folder *genfiles* contains all figures included in the manuscript and the SI Appendix:

- Folder *figures*: contains raw versions of the subplots of Fig. 2, Fig. 3, and Fig. 4 H
- Folder *figures_SI*: contains several figures of the robustness checks, barcharts for Shapley values from all years, and results from the elastic net model (in the folder *LASSO_results*).
- Folder *maps*: contains several maps of the GDP per capita levels for countries and regions, e.g. those in Fig. 4 A-E in the manuscript.

## Replication

The folder *scripts* contains all R scripts necessary to replicate our findings and data. The file 00_master.R calls all other scripts and reproduces the results.
It is necessary to download another folder with the data on famous individuals from here: https://zenodo.org/doi/10.5281/zenodo.11546192.
All files included in the zenodo repository must be saved in a subfolder called *./raw_data/famous_individuals* for the scripts to find it.
Below you find all details on the used packages and R versions:

```
R version 4.2.3 Patched (2023-03-29 r84137)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS 14.5

Matrix products: default
LAPACK: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] doParallel_1.0.17       iterators_1.0.14        foreach_1.5.2           RColorBrewer_1.1-3      treemap_2.4-4           rgdal_1.6-7
 [7] boot_1.3-28.1           sandwich_3.0-2          modelsummary_1.4.1      data.table_1.14.8       BMA_3.18.17             rrcov_1.7-4
[13] inline_0.3.19           robustbase_0.99-0       leaps_3.1               survival_3.5-3          BMS_0.3.5               caret_6.0-94
[19] lattice_0.20-45         ipred_0.9-14            rpart.plot_3.1.1        rpart_4.1.19            dplyr_1.1.4             rsample_1.1.1
[25] ds4psy_0.9.0            forcats_1.0.0           htmlwidgets_1.6.2       pals_1.7                plyr_1.8.8              car_3.1-2
[31] carData_3.0-5           ggpmisc_0.5.3           ggpp_0.5.3              countrycode_1.5.0       fixest_0.11.1           plm_2.6-3
[37] xlsx_0.6.5              plotly_4.10.2           DataVisualizations_1.3.1 raster_3.6-23          sp_2.0-0                sf_1.0-14
[43] eurostat_3.8.2          modi_0.1.2              EconGeo_2.0             cowplot_1.1.1           flextable_0.9.5         officer_0.6.5
[49] kableExtra_1.4.0        knitr_1.46              stringr_1.5.0           broom_1.0.5             ICC_2.4.0               future.apply_1.11.0
[55] future_1.33.0           irr_0.84.1              lpSolve_5.6.20          haven_2.5.3             reshape2_1.4.4          glmnet_4.1-7
[61] Matrix_1.6-1.1          ggplot2_3.5.0

loaded via a namespace (and not attached):
  [1] SparseM_1.81            ModelMetrics_1.2.2.2    maxLik_1.5-2            ragg_1.2.5              tidyr_1.3.0             hardhat_1.3.0
  [7] generics_0.1.3          terra_1.7-39            proxy_0.4-27            tzdb_0.4.0              xml2_1.3.5              lubridate_1.9.2
 [13] httpuv_1.6.11           assertthat_0.2.1        gower_1.0.1             xfun_0.43               hms_1.1.3               rJava_1.0-6
 [19] evaluate_0.21           promises_1.2.0.1        DEoptimR_1.1-1          fansi_1.0.6             readxl_1.4.3            igraph_1.5.1
 [25] DBI_1.2.2               reshape_0.8.9           stats4_4.2.3            purrr_1.0.1             ellipsis_0.3.2          backports_1.4.1
 [31] fontLiberation_0.1.0    insight_0.19.3          gridBase_0.4-7          fontBitstreamVera_0.1.1 vctrs_0.6.5             quantreg_5.96
 [37] here_1.0.1              abind_1.4-5             withr_3.0.0             collapse_1.9.6          bdsmatrix_1.3-6         checkmate_2.2.0
 [43] svglite_2.1.1           pacman_0.5.1            lazyeval_0.2.2          crayon_1.5.2            crul_1.4.0              recipes_1.0.6
 [49] pkgconfig_2.0.3         labeling_0.4.3          units_0.8-2             nlme_3.1-162            nnet_7.3-18             rlang_1.1.3
 [55] globals_0.16.2          lifecycle_1.0.4         MatrixModels_0.5-2      fontquiver_0.2.1        httpcode_0.3.0          dichromat_2.0-0.1
 [61] cellranger_1.1.0        rprojroot_2.0.3         lmtest_0.9-40           datawizard_0.8.0        ISOweek_0.6-2           zoo_1.8-12
 [67] viridisLite_0.4.2       parameters_0.21.1       KernSmooth_2.23-20      pROC_1.18.4             shape_1.4.6             classInt_0.4-9
 [73] parallelly_1.36.0       readr_2.1.5             scales_1.3.0            magrittr_2.0.3          bibtex_0.5.1            compiler_4.2.3
 [79] RefManageR_1.4.0        miscTools_0.6-28        cli_3.6.2               unikn_0.8.0             listenv_0.9.0           Formula_1.2-5
 [85] mgcv_1.8-42             MASS_7.3-58.2           tidyselect_1.2.1        stringi_1.7.12          textshaping_0.3.6       askpass_1.1
 [91] grid_4.2.3              polynom_1.4-1           tools_4.2.3             timechange_0.2.0        rstudioapi_0.15.0       uuid_1.1-0
 [97] tables_0.9.17           prodlim_2023.03.31      farver_2.1.1            digest_0.6.33           shiny_1.7.4.1           lava_1.7.2.1
[103] gfonts_0.2.0            Rcpp_1.0.12             performance_0.10.4      later_1.3.1             httr_1.4.7              gdtools_0.3.7
[109] Rdpack_2.4              colorspace_2.1-0        splines_4.2.3           confintr_1.0.2          xlsxjars_0.6.1          mapproj_1.2.11
[115] systemfonts_1.0.4       xtable_1.8-4            dreamerr_1.2.3          jsonlite_1.8.8          timeDate_4022.108       R6_2.5.1
[121] pillar_1.9.0            htmltools_0.5.5         mime_0.12               glue_1.7.0              fastmap_1.1.1           class_7.3-21
[127] regions_0.1.8           codetools_0.2-19        maps_3.4.1              pcaPP_2.0-3             mvtnorm_1.2-2           furrr_0.3.1
[133] utf8_1.2.4              tibble_3.2.1            numDeriv_2016.8-1.1     curl_5.2.0              zip_2.3.0               openssl_2.1.0
[139] rmarkdown_2.23          munsell_0.5.0           e1071_1.7-13            gtable_0.3.4            bayestestR_0.13.1       rbibutils_2.2.14
```