Assessing future vulnerabilities: a machine learning approach to multi-dimensional population projections at the sub-national level

Andrea Tamburini[1] Erich Striessnig[2] Raya Muttarak[3]

[1]VID, WiC (IIASA, VID/OeAW, Univ. Vienna) [2] Univ. of Vienna, WiC (IIASA, VID/OeAW, Univ. Vienna) [3]Univ. of Bologna, IIASA, WiC (IIASA, OeAW, Univ. Vienna)

Background & Motivation

Importance: spatially-explicit population age structure plays an essential role in **population-environment interactions**:

- **Children and the elderly** are particularly **susceptible to health** stress caused by extreme **heat events and natural hazards**.
- **Age structure** influences **consumption** behavior and **labor** supply which in turn has implications on **environmental impact**.
- The impact of **climate change** varies across **geographical areas**.

This study therefore aims to improve the understanding of spatially-disaggregated **population age structure to inform** better **adaptation and mitigation** planning in the face of global environmental change.

Geographical Setting: The vast demographic and climatic variability in the territory of the EU and the UK provides an interesting test case for the development of a model for sub-national disaggregation of the Shared Socioeconomic Pathways (SSPs).

The **Nomenclature of Territorial Units for Statistics (NUTS)** classification is a **hierarchical system** for dividing up the **economic territory** of the **European Union** (and further) and the **UK** for the purpose of collecting, developing and harmonizing European **regional statistics**, socioeconomic analyses of the regions and framing of **EU regional policies.**

Figure 1. Proportion of elderly population (65+) and mean temperature, European NUTS-2 regions, 2015-19.

Data

- **The Eurostat Database**: contains information on past population distributions by age, sex (and educational attainment) at the level of 249 NUTS-2 regions (complete). Information is reported yearly starting from 1990 until 2021. We focus on six different age groups (under-15, 15-24, 25–44, 45-64, 65+).
- **The Wittgenstein Center Data Explorer**: contains SSP-coherent age, sex (and educational attainment) specific population projections at the national level.
- **Harvard Dataverse**: contains global 1-km^2 population and urban land fractions consistent with the SSPs. Aggregating these data at the NUTS-2 level, we generate total population counts and urban land proportions which are used in the construction of the independent variables.

Methods

An extensive exploratory data analysis informs the development of a **subnational-level** (NUTS-2) model of **five-year changes** in the fractions of **5 age groups** which can then be used for long-term population projections.

Modelling Strategy

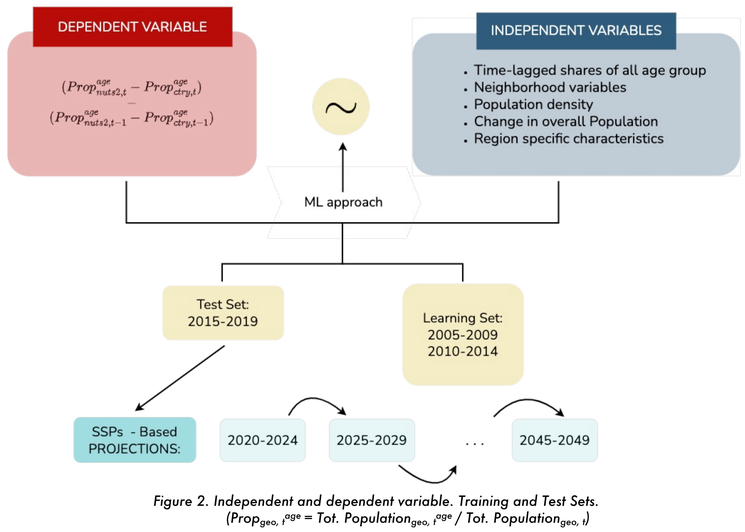


Figure 2. Independent and dependent variable. Training and Test Sets. ($Prop_{geo,\,t}{}^{age}$ = Tot. $Population_{geo,\,t}{}^{age}$ / Tot. $Population_{geo,\,t}$)

Exploratory data analysis reveals complex **non-linear relationships** between **response and predictor variables**. **Tree-based regression** models are capable of integrating these relationships. After testing different possibilities, like Regression Trees, Random Forest and Conditional Inference Trees, the best fit to our data is achieved with a set of **gradient boosted regression trees.** The combined predicted fractions from the five models are then rescaled to sum up to 1 using **iterative proportional fitting** (IPF).

Boosted regression trees combine the **strengths of two** algorithms: **regression trees** (models that relate a response to their predictors by recursive binary splits) and **boosting** (an adaptive method for combining many simple models to give improved predictive performance).

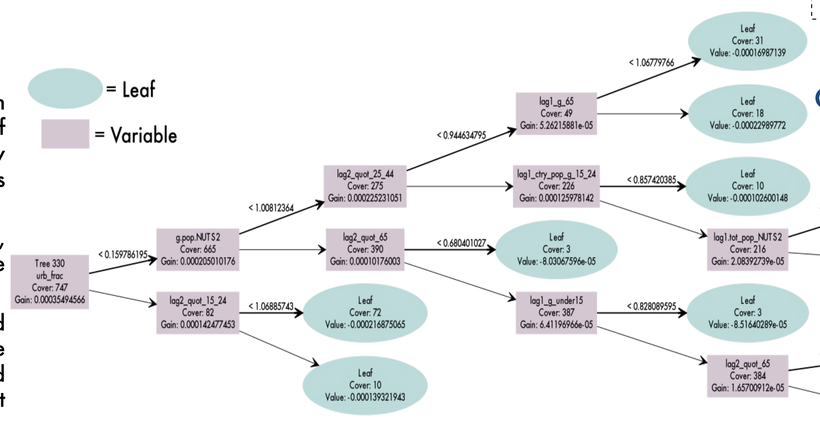


Figure 3. Regression Tree. This is one of the components of the Gradient Boosted Tree.

Validation and Results

Figure 4 shows the results from the **gradient boosted regression tree** for the the age group **65+** for **SSP-2**.

Training Set: data collected for 2005-2009 and 2010-2014 (with relative lagged variables)

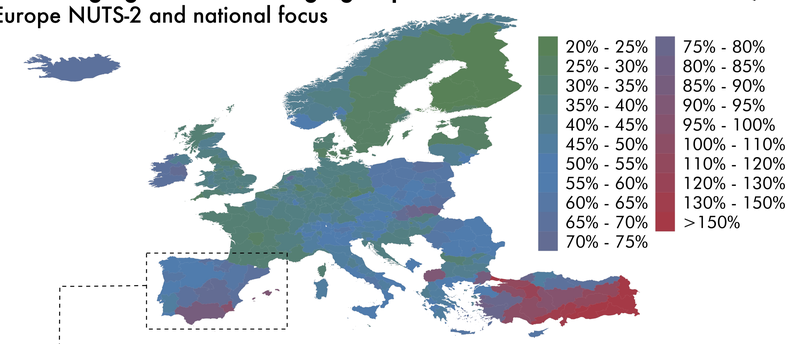
Test Set: 2015-2019 data.

Model Validation: **residuals** analysis, R^2 (above 0.9 also in the testing step) and **comparison** of the **RMSE** with those resulting from a **simple linear regression**, an **AR(1)** model and a **naive model** reproducing the last observed value for each NUTS-2 region.

The Boosted Regression Tree performed better then the competitors.

Main explanatory variables: 1st and 2nd lags in its growth rate, the urban fraction and the lagged total regional population.

Percentage growth for the age group 65+ from 2015-19 to 2045-49, SSP-2
Europe NUTS-2 and national focus



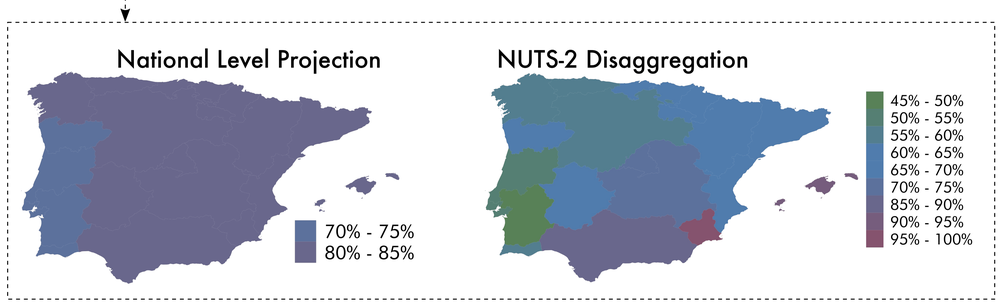


Figure 4. Projection result for the age group 65+. Percentage change of the age group proportion between 2015-19 and 2045-49

Outlook

1- Further improvement of the model relies on the inclusion of additional explanatory variables (number of cities, economic indicators, migration) and their spatial dynamics (spatial random forest, Gaussian processes).

2- Disaggregation of other SSPs to compare the results under different, scenario-based assumptions and to understand their sub-national implications.

3- Further disaggregation by educational attainment on top of the age dimension allowing more detailed climatic risk mapping.

4- SSP disaggregation at the more refined NUTS-3 level to improve spatial accuracy of intervention strategies to reduce climate vulnerability.

Contact: andrea.tamburini@oeaw.ac.at