Distributional Modelling in R

03 - Model Checking - Exercises

In this example we analyze population count data from Austria and Switzerland. The count data originates from the census conducted in Austria and Switzerland during 2021 and 2022.

1. Download the data from:

   https://nikum.org/dmr/Data/AustriaMunicipal.rds
   https://nikum.org/dmr/Data/SwissMunicipal.rds
   https://nikum.org/dmr/Data/AustriaSwissPop.rds

   You can use the following R code

```
R> download_data <- function(data = "AustriaMunicipal.rds") {
+    file <- paste0("https://nikum.org/dmr/Data/", data)
+    tdir <- tempfile()
+    dir.create(tdir)
+    download.file(file, file.path(tdir, data))
+    return(readRDS(file.path(tdir, data)))
+  }
R> AustriaMunicipal <- download_data("AustriaMunicipal.rds")
R> SwissMunicipal <- download_data("SwissMunicipal.rds")
R> AustriaSwissPop <- download_data("AustriaSwissPop.rds")
R> library("sf")
R> plot(AustriaMunicipal)
```

   The data consists of the following variables:

Variable	Description
id	Identification number of the municipality.
country	Country identifier.
area	The area in square kilometers.
pop_km2	Population per square kilometer.
IMD	Sealing density.
NTL	Mean night time light emission.
LC_*	Mean land cover classes.
AIR_*	Airport counts (ct), per square kilometer (km2), mean distance (dist).
CLG_*	College counts (ct), per square kilometer (km2), mean distance (dist).
DOC_*	Doctor counts (ct), per square kilometer (km2), mean distance (dist).
HSP_*	Hospital counts (ct), per square kilometer (km2), mean distance (dist).
MAL_*	Mall counts (ct), per square kilometer (km2), mean distance (dist).
NHO_*	Nursing home counts (ct), per square kilometer (km2), mean distance (dist).
PRK_*	Park counts (ct), per square kilometer (km2), mean distance (dist).
PLG_*	Playground counts (ct), per square kilometer (km2), mean distance (dist).
SCH_*	School counts (ct), per square kilometer (km2), mean distance (dist).
SPM_*	Supermarket counts (ct), per square kilometer (km2), mean distance (dist).
UNI_*	University counts (ct), per square kilometer (km2), mean distance (dist).

2. The objective of this analysis is to develop a model capable of predicting population density in regions lacking observational data. The present dataset serves as a preliminary demonstration, illustrating the potential of integrating census data with satellite imagery and open street map data to generate reliable estimates of population density.

   Hence, the initial step involves identifying an appropriate distribution for the response variable `pop_km2`. Additionally, exploring potential transformations of the response variable is advisable.

3. Next, divide the dataset into separate subsets for Austria and Switzerland for both training and testing purposes. Subsequently, utilize the **gamlss2** package to train a Generalized Additive Models for Location Scale and Shape (GAMLSS) exclusively on the Austrian data. Experiment with different model configurations to identify the most suitable one through iterative trial and error.

4. After selecting the top three models, compute the randomized quantile residuals using the Swiss dataset and compare their performances.

5. Additionally, calculate the Continuous Ranked Probability Score (CRPS) for each model to determine which one exhibits the best overall performance.

6. Finally, generate a comprehensive probabilistic forecast for population density in Switzerland. This entails computing the mean, median, as well as the 5% and 95% quantiles. Evaluate the predictive accuracy of these forecasts using the mean squared error for the mean and median estimates, and the pinball loss for quantiles, utilizing the `pinLoss()` function from the **qgam** package.

7. Visualize predictions using the **sf** package.