Applied Statistics and Statistical Software (02441)
Technical University of Denmark
Department for Applied Mathematics and Computer Science

HTK Case: Energy performance of buildings

Background

Høje Taastrup municipality (HTK) is interested in lowering their energy consumption in their public buildings. They cannot afford just retrofitting all their buildings. They want to use statistical tools to identify those buildings that have the poorest performance in order to prioritize those buildings first.

For older buildings a large proportion of the total energy consumption is used for heating. From physics it is known that the heat loss through a simple wall is given by:

$$Q_{heat} = U_A(T_{indoor} - T_{outdoor})$$

where U_A is a measure of the amount of insulation. If it is assumed that the indoor temperature is relatively constant the the effective average U_A for the entire building can be estimated using regression methods.

Other climatic variables and patterns in the use of buildings may also influence the heat loss.

About the data

To do the statistical modelling you are provided with daily readings for the district heating meters in 97 buildings in HTK. There is one file per day and the file name includes a timestamp for the time of the reading. All these files are included in `meterdata.zip`.

You also get almost hourly climate data from WUnderground in the file `WUndergroundHourly.RData`. These data are downloaded from the WUnderground API (Using 'library(rwunderground)') with Lat=55.65, Long=12.28 and the 'history' function.

Assignment

1. Read and merge the two data sources (**Data Cleansing part**)
   - WUnderground data
     (a) Read the `RData` file
     (b) Exclude columns with pure `NA`s or fixed values
     (c) For each day you should calculate
         – the mean value for continuous variables
         – the mode of the factor variables
     (d) Aggregate your results into a single dataframe
   - Meter data
     (a) Get an overview over the file structure in `meterdata.zip`
     (b) Read all data into a single dataframe
     (c) Keep only columns 1, 2 and 4 ("ID", "Time" and "Reading")
     (d) Some meters have problems with the readings. Exclude meters with less than 121 records to avoid long gaps

HTK Case: Energy performance of buildings

(e) For each building you should calculate the mean consumption (one number per building per day). The consumption can be understood as the difference between daily readings. The problem is that readings were not taken at the same time points for all days. To correct for this use the daily readings to interpolate a reading at 11.59pm for each day.

(f) Merge the meter and the WUnderground dataframes

(g) Include summary() of your merged dataframe in your report and compare with the version that is handed out

(h) Include the number of rows in your data.frame and the number of remaining meters in your report

(i) Include your code for merging the data as an appendix

2. Perform appropriate statistical analysis to identify buildings which could most benefit from a retrofit (**Analysis Part**)

   (a) Load data `merged_data.csv` (from CampusNet). Do NOT use your own merged data!

   (b) Present the data and get an overview

   (c) Make appropriate linear regression model(s)

   (d) Present the model(s) with focus on energy performance of the buildings

   (e) Comment on your findings - this may include recommendations to HTK

   (f) **BONUS**: Use extra information about the buildings (`HTK_building_data_share.xlsx`) for interpretation of your results.

   (g) **HINT**: Start by establishing a simpler model(s). Assume $T_{indoor} = 21C$ and establish a model(s) which only includes Date, I(21-temp) and ID. At a later stage you can increase complexity of your analysis and also include weather variables.

The following functions may come in handy: `dir`, `as.POSIXct`, `aggregate`, `table`, `approx`, `diff`, `merge`, `rbind`