Independent work 2 – assessing data-related uncertainties and models' predictive performance

Data: air temperatures across Finland ("*AirTemperatureData.csv*").

Uncertainty. It is important to acknowledge that all statistical models are products of their underlying data. In real-life it is impossible to measure everything; for example location and properties of every plant individual in the Fennoscandia (constituting the *population*). Thus in practice we are limited to model smaller samples of the whole population. This means that the outcome of our modelling *is likely to be* dependent on the data sample. We might ask: would the results be significantly different if another data sample is drawn from the population (the basis of statistical significance testing), or how large are the uncertainties in the statistical measures that we calculate from the data sample?

Bootstrapping (i.e. repeated random sampling with replacement) is a highly efficient way to assess data sample-related uncertainties in statistical analyses. Using bootstrapping we get an idea how well a specific sample data, or derived statistical measures, are representative of the population.

Question 1: Using bootstrap sampling with 1000 repeats, calculate sample mean and associated 95 % confidence intervals for mean air temperature (as in week 1 "*day2_exercises.pdf*"). Plot the result as a histogram with confidence intervals indicated as dashed lines. Save the figure.

Question 2: Model air temperatures using first order polynomial terms of elevation and sea as predictors. By the means of bootstrap-sampling (1000 repeats), quantify the 95 % confidence interval for estimated regression slope for elevation. Plot the results as a histogram with confidence intervals indicated as dashed lines. Save the figure. You may use and further modify the function below:

```
b <- function(){
        sam <- sample(nrow(d), replace = TRUE)  # draw a bootstrap sample
        m <- lm(y~x,data=d[sam,]) # fit a linear regression model
        return(coef(m)) # return estimated coefficients. Use "[]" –syntax to extract specific model
                        # coefficient
}
```

Predictive performance. Statistical models can be used to predict the value of a response variable, when predictor values are known. This is valuable since (i) it allows for a creation of spatially continuous data (maps) to support decision making, (ii) increase the scientific understanding of how a particular phenomenon varies in nature, and (iii) is often only source of information for remote areas that are difficult to access. While performing a model prediction is fairly straightforward, the prediction is "useless" if a scientist has no information about the goodness of the prediction. In practice the only way to assess prediction performance of a model is via cross-

validation. This means that prior to model fitting, some of the data points are set to side that are not used for constructing the model (so called evaluation data). Then the fitted model can be used to predict the response variable's values over the evaluation data, allowing a comparison between observed and predicted values. Common measures of predictive performance are e.g. mean difference, root mean squared error (RMSE) and correlation between the observed and predicted values (r).

A special case of cross-validation is "leave-one-out cross-validation" (LOOCV). This means that cross-validation is performed as many times as there are observation points; at each cross-validation round, the model is fitted using *n*-1 observations and then predicted to the one remaining data point (i.e. the one that was not used for model fitting). This procedure is repeated until all data points have been set aside once. After running LOOCV you have observed and predicted value for each data point thus enabling a comparison.

Question 3: Write a *for*-loop to perform a leave-one-out cross-validation of a model, where air temperatures are being predicted using elevation (first and second order polynomial terms). Create a scatterplot, where observed air temperatures are on y-axis and predicted air temperatures on x-axis. Quantify the agreement between observed and predicted values by the means of mean difference and Pearson's correlation coefficient. Add both measures to the scatterplot.

Question 4: Does the predictive performance increase after including latitude (y) and longitude (x) to the previous model? Consider their first and second order polynomial terms & and their interaction. Create a scatterplot, where observed air temperatures are on y-axis and predicted air temperatures on x-axis. Quantify the agreement between observed and predicted values by the means of mean difference and Pearson's correlation coefficient. Add both measures to the scatterplot.

Tips for LOOCV:

- to define the for-loop, you need to know the number of rows in the data; this can be obtained using a function *nrow()*
- at each iteration (=loop-round), you need to set aside one row of the whole data for evaluation in turn; other are used for fitting the model
- you need to collect the predicted values of each iteration round to a result vector; before initiating the for-loop, create empty vector for this purpose
- inside the loop use *c()* –function to collect the predicted values to the result vector

Return your answers via Moodle as a single .pdf –file **by latest Wednesday 15.11.2017**. Please name the file as "*Lastname_Week2_Assigment.pdf*". Include any relevant figures and R-script to the file.