Math-710 Data analysis for science and engineering
Spring 2017

Proposal for the final project

Submitted by Franziska Meinherz

Preliminary note:
Due to a slight reorientation in my PhD (I am still in the 1st year), I am now at a point where I need to do a lot of conceptual work. I do not currently have research questions requiring statistical analysis. Those will emerge out of the conceptual work in which I am currently engaged. Therefore, for this semester project, I would like to work on a very pressing problem from an association of which I am part (Octanis Association - Rover project). I am confident that applying the methods and concepts which we have learnt in class to this problem will enhance my practical understanding of them such that I can then easily apply them to my PhD, once I am more advanced in it.

Research problem

Problem context

Octanis Association has built a test rover which can perform data gathering tasks in difficult terrain, and which can be built with basic and low-cost components and runs with open source software. At the moment, the test rover has the sensors required for gathering data on the structure of the snow cover of any given area, and has been sent on a mission to Antarctica undertaken by researchers from the University of Grenoble. The rover is powered by solar cells which are mounted on its top and has a battery to store energy. The data which it gathers is stored on an SD card which is built into the rover. The rover both gathers data regarding its task - in this case, the structure of the snow cover - and on its own performance.

The rover will be back from its current mission towards the end of May. As a follow-up to this mission, the current rover will be improved, in order to enhance its usefulness for similar missions. One key challenge is to enhance the time during which the rover can be active, by optimising the process of charging and discharging the battery.

Problem statement

The aim of this research project is to use the data which the rover has gathered on its battery and battery charging / discharging status, as well as on its context, to develop an optimising procedure for when the rover should rest and charge its batteries, and for when it should be

active and use energy, in order to maximise its operation time. The project thus needs to address the following points:

1. Data extraction and structuring: The data collected by the rover is stored in ROS .bag files, is ordered by different topics, and can be accessed through command line tools which are comprised in the ROS package (see additional details under "Data"). The data thus has to be extracted into .csv files and organised and structured in a way that facilitates its manipulation with R and which corresponds to the purposes of this semester project.

2. Data exploration: Since the current rover is a pilot project, it seems appropriate to develop tools to visualise the data in ways which allow to identify potential measuring or storing errors, and which additionally provide an overall picture of the rover's functioning, in order to develop first hypotheses as to how its operation time could be enhanced.

3. Data analysis: Data on the rover's battery charge status and battery charging / discharging has to be compared to data on the current UV index and other indicators for battery charging potential, on wind intensity and direction in comparison to the rover's direction of movement, outside temperature, and the rover's speed, in order to find to an optimising strategy for the conditions under which the rover should rest and charge its battery, and for when it should move and use energy.

Data

Since the arrival time of the rover and the data it has gathered during its mission is scheduled only for the end of May, for this project, data from test runs undertaken in Renens will be used. The data structure is exactly the same between the data gathered during the test runs in Renens and the actual mission in Antarctica. Therefore, it should not matter which is used as an input to the data analysis script which is to be developed in the course of this project, with the exeption that no pertinent conclusions can be drawn from the test runs, since the setting was artificial and the sample size - in this case the amount of time which the rover spent in activity - is too small.

Data format and structure

The data is stored in .bag files. This file structure has been developed as part of the ROS (Robot Operating System) environment to store data collected by any robot as soon as it is received. The data is thus organised along a timeline which corresponds to the robot's internal clock (although this can be overridden with a simulated clock). In this process, the data is organised in different topics, which each can store different variables. The data can be accessed through different command line tools which are available in the ROS package and can be played back as text in the command window. The data is accessed by topic. The ROS package also provides some tools to visualise the played back data.

Variables

For this project, two kinds of variables will be considered: On the one hand, variables which relate to the rover's functioning, such as its battery status, its battery charging / discharging status, its speed, its energy use per achieved motion, and its direction, and on the other hand, variables which describe the rover's surroundings and the situational context, such as variables describing the available solar power, variables on wind intensity and direction, information on the structure of the terrain - both on the macro level (elevation and inclination) and the micro level (snow cover structure), and the current air pressure.

Methods

Data extraction and formatting

1. Data extraction: böh?

2. Data structuring: The .csv files obtained in the previous step are to be organised in such a way that the data is accessible in ways which makes sense given the steps which are to be performed in the data analysis part. As the project advances, the decisions regarding the formatting of the data are thus revisited in an iterative process.

Data exploration

1. Finding measuring / storing errors: Data must be checked for invalid entry points as well as for unreasonable outliers. For the first, böh?!?!. For the second, the data is to be visualised in box plots. In cases where realistic data ranges cannot easily be determined, the distribution as shown in the box plots is used as an indicator for what can be assumed a realistic data range, and outliers to this derived realistic range are then scrutinised for their likelihood to be erronic values. Depending on the variable, this can be done by looking at the values recorded shortly before and after, or by cross-comparing the outliers by looking at their relative similarity as well as at the values of the other variables recorded at the same time amongst similar outliers. In cases where realistic data ranges can easily be determined, outliers lying outside of them are to be eliminated.

2. Providing an overall picture: The aim of this step is to draw a picture of the rover's functioning depending on the situational context. Therefore, the variables describing the rover's battery status, and battery charging / discharging are to be plotted against variables describing the situational context (weather data etc.) as well as against variables describing the rover's activity (speed, direction, elevation gain / loss etc.)

Data analysis

Different analytical tools are to be used in order to gain a picture which is as consistent as possible with what is actually influencing the rover's battery lifetime.

1. Correlation analysis: Variables describing the rover's functioning are constrasted with variables describing the situational context in a cross-correlation analysis, in order to identify on the one hand whether there seem to be patterns in how they co-develop, and on the other hand, to identify the situational variables which most seem to influence the rover's functioning.

2. Factor analysis/Principal component analysis: The aim of this exercise is to see whether there are clusters emerging describing specific modes of functioning in specific circumstances. Such clusters could in a next step be used to identify optimal strategies for the rover. This exercise is to be done both through a principal component analysis, where the emerging clusters are then analysed regarding their logical meaning and consistency, and through a factor analysis, where the factors are predefined based on the insights from the correlation analysis and hypotheses regarding the logic underlying the rover's functioning, and then tested regarding their explanatory power, internal consistency, and significance.