University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Martin Leimer

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, April 16, 2012

Bachelorarbeit in Informatik
Datenbanktechnologie

Topic: Statistical Comparison of Regions in the Swiss Feed Database

The current on-line interface to query feed data linked to geographic information is so far limited to the selection criteria canton and altitude. Feedback from potential end-users show the need to extend the filter criteria to allow queries on local feed data that can be compared with other regions of similar altitude and the national average.

The goal of this thesis is to study and apply the F-Test and T-Test for the statistical comparison of regions based on the containment of the nutrients that are found in feed samples of these regions. The procedure involves several steps. In a first step, the F-test is applied to test for equality/inequality of variance of the two populations. Depending on this result, either the formula for equal or unequal variances must be chosen in the subsequent Student's t-test. In general, unequal sample size must be assumed. The t-test suits for univariate problems. A generalization of Student's t statistics, called Hotelling's T-square statistic, allows for the testing of hypotheses on multiple (multivariate), often correlated, measures, which is characteristic for feed data. Both, F-test and t-test are based on the assumption of a normal distribution. This may not always be the case, particularly with respect to minerals and trace elements. Normality can be tested by using the Shapiro-Wilk or Kolmogorov-Smirnov test.

The thesis pursues the following outcome functionality. First, the user selects two locations on the map using a mouse pointing device and, then, the system automatically gather all feed samples that are found in the surroundings of these two locations and computes the tests. The result is displayed using a table that also incorporates other relevant statistics to the locations as min, max and averages of the containment of the selected nutrients. The second type of query aims for the selected regions to find out the top-k most similar(dissimilar) regions based on the results of the two tests.

This thesis is to be completed in close collaboration with research authorities of Agroscope, including one day visit to the agriculture research institute in Posieux.

Deliverables:

1. Implementation and integration of the F-Test and T-Test into on-line application of the Swiss Feed Database.
2. Bachelor thesis presenting your results.
3. Presentation of the results (15 minutes).

Literature:

- Lozan, José and Hartmut Kausch. Angewandte Statistik für Naturwissenschaftler. 4th ed. Hamburg: Wissenschaftliche Auswertungen, 2007.
- Hartung, Joachim, Bärbel Elpelt and Karl-Heinz Klösener. Statistik. 12 th ed. München: Oldenbourg, 1999.
- Abramowitz, Milton and Irene A. Stegun. Handbook of Mathematical Functions With Formulas, Graphs, and mathematical Tables. 10th printing. New York: Wiley, 1972.

Supervisor:

- Andrej Taliun

Starting date: 16.04.2012

Ending date: 16.08.2012

Department of Informatics, University of Zurich

Prof. Dr. Michael Böhlen