

Report – Queries on local feed quality – F-Test

1. Problem

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. This approach can be made by applying the F-Test based on the containment on the nutrients that are found in feed samples of these regions.

2. F-Test

2.1. Cumulative F-Distribution

With the cumulative F-Distribution function, we can calculate to which level of significance the null hypothesis¹ can be withdrawn or not, given the F-Value and degrees of freedom. The following formula tells to which probability the null hypothesis should be maintained.

$$\int_x^{\infty} f(x|m; n) dx = \int_x^{\infty} m^{\frac{m}{2}} * n^{\frac{n}{2}} * \frac{\Gamma\left(\frac{m}{2} + \frac{n}{2}\right)}{\Gamma\left(\frac{m}{2}\right) * \Gamma\left(\frac{n}{2}\right)} * \frac{x^{\frac{m}{2}-1}}{(mx + n)^{\frac{m+n}{2}}} dx$$

- x F-Value
- m First degree of freedom (FG_{between})
- n Second degree of freedom (FG_{within})
- $\Gamma(x)$ Gamma-function² at x

2.2. F-Value

The F-Value value can be obtained by dividing the mean square value between the groups through the mean square value within the groups:

$$F = \frac{MAQ_{\text{between}}}{MAQ_{\text{within}}}$$

2.2.1. Mean square value (MAQ)

The mean square value can be computed by dividing the corresponding sum of deviation squares through the corresponding degree of freedom.

$$MAQ_i = \frac{SAQ_i}{FG_i}$$

¹ H_0 : The differences between the spot tests are stochastically. – If H_0 is rejected in favour of H_1 , the alternative hypothesis, it is equivalent to say that the variability between the spot tests is significantly different than within at least one spot test.

² The calculation including a gamma function may be very time consuming, especially on pocket calculators. For better performance, the following approach can decrease the execution time dramatically:

- for $x > 0$: $\Gamma(x + 1) = x * \Gamma(x)$
- if x is an integer: $\Gamma(x) = (x - 1)!$

Since the gamma function must only be calculated for integers and multiples of 0.5 (m and n are integers), it is sufficient to know the value of the gamma function for $\Gamma(0.5) = 1.77245$. All other needed values can then be derived with these formulas.

Whereas the degree of freedom is calculated as follows:

$$FG_{\text{between}} = k - 1$$

$$FG_{\text{within}} = n - k$$

Given these formulas, it can be implied that:

$$MAQ_{\text{between}} = \frac{SAQ_{\text{between}}}{k - 1}$$

$$MAQ_{\text{within}} = \frac{SAQ_{\text{within}}}{n - k}$$

n Total number of all measurement values

k Number of spot tests / groups

2.2.2. Sum of deviation squares (SAQ)

The sum of deviation squares must be calculated for the between and within the groups variability. For control reasons it may be helpful to calculate the overall variability as well, since the sum of the between and within the groups variability must equal to the overall variability!

Overall variability (SAQ)

$$SAQ_{\text{overall}} = \sum (x_{ij} - \bar{x})^2 = \sum x_{ij}^2 - \frac{(\sum x_{ij})^2}{n}$$

Between the groups (SAQ)

$$SAQ_{\text{between}} = \sum n_j * (x_j - \bar{x})^2 = \sum \left[\frac{(\sum x_{ij})^2}{n_i} \right] - \frac{(\sum x_{ij})^2}{n}$$

Within the groups (SAQ)

$$SAQ_{\text{within}} = \sum \left[\sum (x_{ij} - \bar{x}_j)^2 \right] = \sum x_{ij}^2 - \sum \left[\frac{(\sum x_{ij})^2}{n_i} \right]$$

n total number of samples over all groups

n_i number of samples in the i^{th} group

x_{ij} j^{th} sample of i^{th} group

\bar{x} arithmetic average

2.3. Validity and restrictions

The F-Test is only appropriate in an environment where we have at least two spot tests, each including several samples. Moreover, the spot tests must be non-affiliated and should be normally distributed, especially if there are only two spot tests. However, if there are more than two spot tests, the more the spot tests contains the same number of samples, the more significant meanderings to the normal distribution in each spot test is allowed.

On the other hand, there is neither a restriction on the quantity in a spot test nor a restriction on the quantity between several spot tests.

3. Analysis

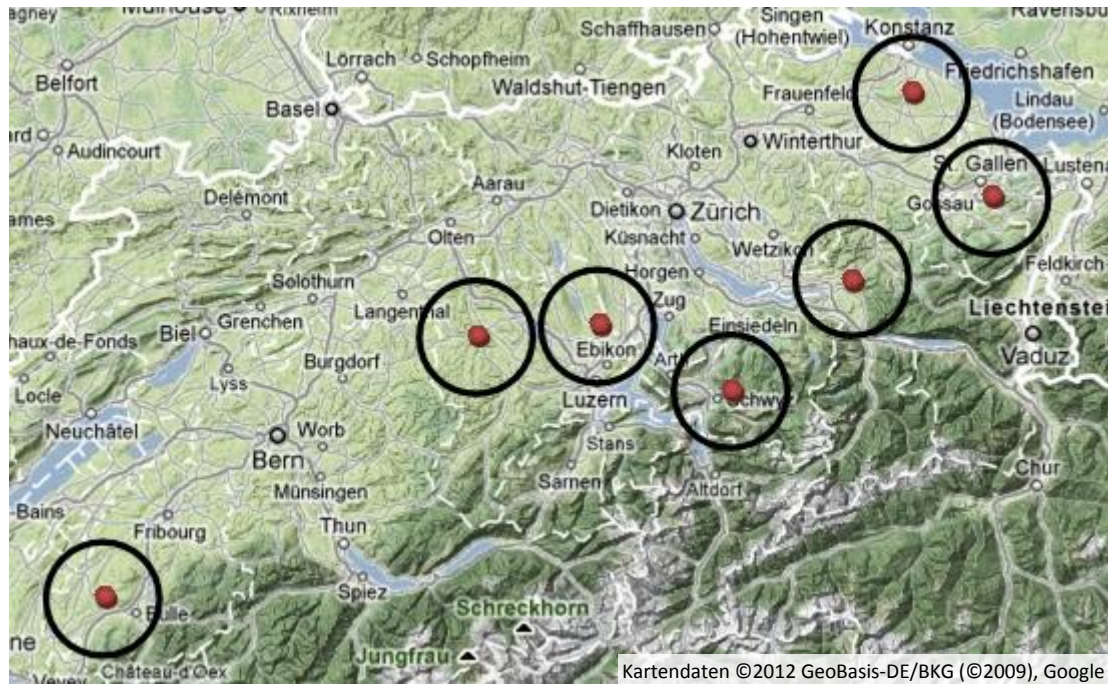


Figure 1: Selected regions for F-Test comparison

The analysis was made on the nutrients phosphor and net energy growth for seven locations including samples within a range of 10km each as seen in Figure 1. The locations were taken such that there are geographical similar regions, enough samples to evaluate and a good sample distribution in all cardinal points within the spot tests.

The goal of the analysis was to identify if there exists a correlation in the measurements between geographical similar regions or if other factors have a greater impact.

3.1. Phosphor

| | Sales | Berg | Willisau | Schwyz | Eschenbach | St. Gallen | Ernetschwil |
|-------------|---------|---------|----------|---------|------------|------------|-------------|
| Sales | 100.00% | 52.67% | 27.91% | 39.44% | 4.30% | 0.62% | 0.83% |
| Berg | 52.67% | 100.00% | 44.36% | 12.16% | 3.56% | 0.40% | 0.00% |
| Willisau | 27.91% | 44.36% | 100.00% | 4.80% | 12.41% | 2.37% | 0.00% |
| Schwyz | 39.44% | 12.16% | 4.80% | 100.00% | 0.56% | 0.19% | 39.25% |
| Eschenbach | 4.30% | 3.56% | 12.41% | 0.56% | 100.00% | 32.99% | 0.00% |
| St. Gallen | 0.62% | 0.40% | 2.37% | 0.19% | 32.99% | 100.00% | 0.00% |
| Ernetschwil | 0.83% | 0.00% | 0.00% | 39.25% | 0.00% | 0.00% | 100.00% |

For the nutrient phosphor, we can see many regions where the null hypothesis should not be withdrawn to a level of significance of 5%. Unfortunately, there seems to be no direct geographical correlation. For example Berg is near the Lake Constance, but this region is more similar to Willisau in Emmental wide away from a lake than to Eschenbach near the Lake Lucerne. On the other hand, there seems to be some correlation between Ernetschwil and Schwyz. Both locations contain samples which were gathered in average higher in altitude than the other ones. Furthermore, both centres have nearly the same distance to a lake.

3.2. NEV (Net energy growth)

| | Willisau | Eschenbach | Sales | Schwyz | Ernetschwil | Berg | St. Gallen |
|-------------|----------|------------|---------|---------|-------------|---------|------------|
| Willisau | 100.00% | 0.69% | 0.32% | 0.17% | 0.00% | 0.00% | 0.00% |
| Eschenbach | 0.69% | 100.00% | 36.50% | 0.00% | 0.00% | 0.00% | 0.00% |
| Sales | 0.32% | 36.50% | 100.00% | 35.53% | 0.31% | 0.03% | 0.00% |
| Schwyz | 0.17% | 0.00% | 35.53% | 100.00% | 12.86% | 1.53% | 0.00% |
| Ernetschwil | 0.00% | 0.00% | 0.31% | 12.86% | 100.00% | 50.43% | 0.14% |
| Berg | 0.00% | 0.00% | 0.03% | 1.53% | 50.43% | 100.00% | 0.23% |
| St. Gallen | 0.00% | 0.00% | 0.00% | 0.00% | 0.14% | 0.23% | 100.00% |

For the net energy growth the null hypothesis should be withdrawn for most of the regions. There were the null hypothesis should not be withdrawn to a level of significance of 5%, no geographical similarities can be found at a glance. Although Sâles, Eschenbach and Ernetschwil are all near a lake and seems to have some similarity according to the F-Test, for Berg there cannot be found any, although it has more or less the same geographical properties as Eschenbach.

4. User-Interface

After the query has been executed and the results are shown to the user, he can select “region-comparison” in the drop-down list. Then, he first has to choose a radius, which defines how many nearby feed samples will be gathered into one group. In a second step, he has to identify two locations, which will be taken as the centre of each group, by clicking the corresponding button before selecting a given sample on the map. For usability purposes, the coordinates of the centre will be shown as well as a circle with the selected radius around the chosen centre.

In a last step, the user can now run the F-Test by clicking on the “Calculate F-Test”-Button.

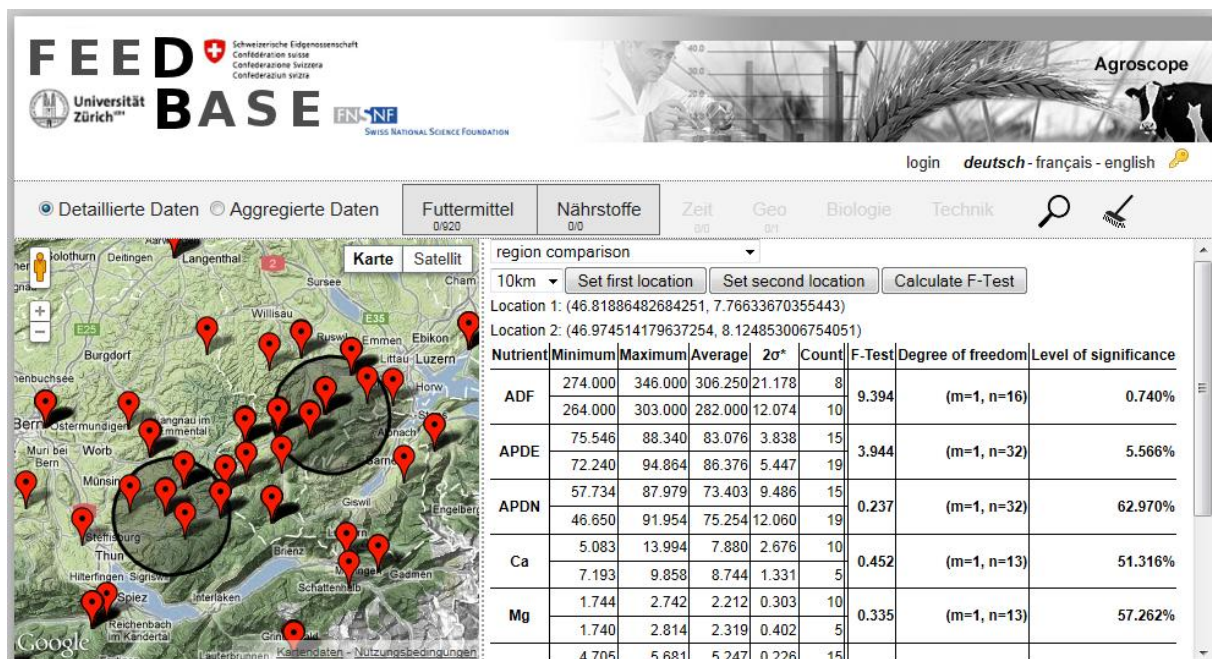


Figure 2: User-Interface for F-Test region comparison

Within a few seconds, the F-Value and the level of significance is calculated for every nutrient and displayed in a table, as seen in Figure 2. Furthermore, the total number, minimum, average, maximum and standard deviation of each nutrient and spot test is shown as well. It is now up to the user to compare other regions as well. For this, the user can change a location and even the radius, by clicking the corresponding buttons and values in the drop-down list.³

5. Implementation

The implementation had been made after having obtained a deep understanding about the F-Test. First of all, the new block element "region comparison" had been made and linked in the drop-down list. Afterwards, the buttons to select a region were added, synchronous to the set-up of the map-interface. If the user clicks on a sample, the *store_selected_location(latlng)*-function is invoked.

```
function store_selected_location(latlng) {  
  
    if(store_next_selected_location > -1) { // every time a sample on the map is cli  
  
        selected_locations[store_next_selected_location] = latlng;  
  
        if(store_next_selected_location == 0) {  
  
            if(circle_location_1 != null) circle_location_1.setMap(null);  
            circle_location_1 = new google.maps.Circle({  
                center: latlng,  
                radius: radius_for_region_comparison * 1000,  
                map: g_map  
            });  
            g_map.fitBounds(circle_location_1.getBounds());  
  
            radius_for_region_comparison_loc_1 = radius_for_region_comparison;  
            $("#first_location_value").html("Location 1: " + latlng);  
  
        } else if(store_next_selected_location == 1) {  
  
            if(circle_location_2 != null) circle_location_2.setMap(null);  
            circle_location_2 = new google.maps.Circle({  
                center: latlng,  
                radius: radius_for_region_comparison * 1000,  
                map: g_map  
            });  
            g_map.fitBounds(circle_location_2.getBounds());  
  
            radius_for_region_comparison_loc_2 = radius_for_region_comparison;  
            $("#second_location_value").html("Location 2: " + latlng);  
  
        }  
  
        store_next_selected_location = -1; // location set.  
  
        if(selected_locations[0] != null && selected_locations[1] != null) { // if  
            document.getElementById("calculate_f_test").removeAttribute("disabled");  
        }  
    }  
}
```

Figure 3: Code example 1: *store_selected_location(latlng)*-function

³ A radius change will not take effect for any location, as long as the location is not redefined afterwards. Please note, that it is possible to have two locations with different radius!

Because the location should not be changed unless the user has clicked the corresponding “Set # location”-button right before, a tracking variable had to be implemented, since this method is invoked every time the user clicks on a map, no matter if the button was clicked before or not. The tracking variable *store_next_selected_location* is set to the location number which will be defined after the user has clicked the corresponding button. If no button was clicked, or the location has already been set, the value is set to -1 such that the storing process is not executed. Then, using AJAX, the parameters of the initial query are complemented with the radius-restriction of every spot test individually. After the two modified queries are executed, the results are given back to the JavaScript-File, where the final calculations are made as seen in Figure 4.

```
//F-Test
var count = new Number(secondArray[5]) + new Number(secondArray[13]);
var sumOfX = new Number(secondArray[6]) + new Number(secondArray[14]);
var sumOfX2 = new Number(secondArray[7]) + new Number(secondArray[15]);
var sumOfXsumOfXByN = new Number(secondArray[8]) + new Number(secondArray[16]);

var betweenTheGroupVariability = (sumOfXsumOfXByN - (sumOfX*sumOfX/count)).toFixed(3);
var withinTheGroupVariability = (sumOfX2 - sumOfXsumOfXByN).toFixed(3);

var MAQbetweenTheGroup = betweenTheGroupVariability;
var MAQwithinTheGroup = withinTheGroupVariability/(count-2);
var Ftest = (MAQbetweenTheGroup/MAQwithinTheGroup).toFixed(3);

var levelOfSignificance;
var degreeOfFreedom = "(m=1, n="+count-2)";

if(new Number(secondArray[5]) == 0 || new Number(secondArray[13]) == 0) {
    Ftest = "---";
    degreeOfFreedom = "---";
    levelOfSignificance = "---";
} else {

    var ft=ftail(Ftest,1,(count-2));
    levelOfSignificance = new Number(RoundDP(ft,5));
    levelOfSignificance *= 100;
    levelOfSignificance = levelOfSignificance.toFixed(3);

}
```

Figure 4: Code example 2: Final calculation steps

The difficulty was to modify the query appropriate and get the right results back. As the corresponding final calculations to get the F-Value were relatively easy, the calculation of the level of significance was difficult again, due to the integral and the gamma function.

6. Reference list

- Lozan, José and Hartmut Kausch. *Angewandte Statistik für Naturwissenschaftler*. 4th ed. Hamburg: Wissenschaftliche Auswertungen, 2007.
- Hartung, Joachim, Bärbel Elpelt and Karl-Heinz Klösener. *Statistik*. 12th 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.
- Andress, Hans-Jürgen. *F-Verteilung*. September, 10th 2001. April, 5th 2012. <<http://eswf.uni-koeln.de/glossar/surfstat/fvert.htm>>.