How to do a nearest neighbor analysis using CANOCO and Excel

1. Using a word processing program, add the physiognomic scores for the site(s) to be analyzed to the CANOCO input file which contains the physiognomic scores for the CLAMP 3A dataset (this file is available from J.A. Wolfe). Save the new file as a text file.
2. Perform correspondence analysis of this new datafile using CANOCO.
   a. Open CANOCO
   b. When asked, specify the new datafile as the "species data" file
   c. When asked, specify correspondence analysis (CA) as the analysis type
   d. When asked, request output of sample scores.
3. Open the CANOCO output file in Word and modify the format from space delimeted to tab delimited so the data can be imported into Excel.
   a. Erase the header information that preceeds the sample scores.
   b. Erase the sample heterogenity and squared residual length scores that follow the sample scores.
   c. Use the global find and replace function to change all spaces to tabs.
4. Copy and paste the sample scores into Excel.
5. The first five columns in Excel will be: N = sample number, NAME = name of site, and AX1, AX2, and AX3 = axis scores. Erase the other columns [1];.
6. Divide the scores in the AX 1, AX2, and AX 3 columns by 100 (the scores are reported x 100 in the CANOCO output file).
7. Add the mean annual temperature (MAT) for the CLAMP sites to the file (this data available from J. A. Wolfe) and for the new site(s), if available.
8. Go to the row for the site which is to be analyzed, and calculate the Euclidean distance after equation 5 from Stranks and England (1997)[2]:

$$E_{jk} = \sqrt{\sum_{h=1}^{D} (X_{hj} - X_{hk})^2} \qquad (1)$$

   Where E_{jk} is the Euclidean distance betwen sites *j* and *k*, $(X_{hj} - X_{hk})^2$ is the distance between sites *j* and *k* along dimesion (or axis) *h*, and *D* is the number of dimernsions (orthogonal axes) that contain useful information.
   a. Insert three new columns and call them x1, x2, and x3.
   b. For column x1, write the equation: (value of AX1 for site to be analyzed) - (cell address for AX1 score). For x2, (value of AX2 for site to be analyzed) - (cell address for AX2 score) and x3, (value of AX3 for site to be analyzed) - (cell address for AX3 score) (Table 1). The resulting values should be equal to 0.
   c. Add a new column after x3 and call it ED (for Euclidean distance). Write the equation: = SQRT(cell address for x1 value^2+cell address for x2 value^2+cell address for x3 value^2) (Table 1).
9. Copy these equations and paste them so that the Euclidean distance is calculated for all the sites.

---

[1] Stanks and England (1997) obtained the most accurate estimates of mean annual temperature for a dataset of 32 sites from New Zealand using three dimensions (i.e. the first three orthogonal axes in CA) and 20 nearest neighbors. Thus, these parameters are used in this handout, though the reseracher should feel free to experiment with different numbers of dimensions and nearest neighbors.

[2] Stranks, L., and England, P., 1997, The use of a resemblance function in the measurement of climatic parameters from the physiognomy of woody dicotyledons: Palaeogeography, Palaeoclimatology, Palaeoecology, v. 131, p. 15-28.

10. Now copy the entire dataset, and do a paste special: values. This will allow the data to be sorted more easily.
11. Sort the database, with the Euclidian Distance (ED) column as the key (ascending). The site to be analyzed will now be at the top of the worsheet, and the rest of the sites will follow in the order of closest to furthest away (in terms of Euclidean Distance) (Table 2).
12. The MAT of the chosen site, T_c , can now be calculated using equation 6 of Stanks and England (1997):

$$T_c = T_0 + a_1x_1 + a_2x_2 \ldots\ldots a_ix_i \qquad (2)$$

where x_i is the distance of that site from the site under analysis along the *i*th axis, or in other words, x1, x2...xi of the Excel spreadsheet. The coefficients a_i and T_0 can be calculated by performing a multiple regression analysis of the data for the 20 nearest neighbors[3], with the MAT as the dependent variable, and x1, x2, and x3 as the independent variables, as illustrated in Table 3. For the chosen site, the value of x1, x2, and x3 is zero, so the MAT estimate is equal to the regression constant.

Table 3: Multiple regression analysis for the San Jose sample, using Data Desk 3.0

Dependent variable is MAT
R^2 = 63.1% R^2(adjusted) = 56.1%
s = 2.399 with 20 - 4 = 16 degrees of freedom

Source	**Sum of Squares**	**df**	**Mean Square**	**F-ratio**
Regression	157.276	3	52.4	9.11
Residual	92.0963	16	5.75602	
Variable	**Coefficient**	**s.e. of Coeff**	**t-ratio**	
Constant	21.2793	0.6939	30.7	
x1	-6.86648	1.504	-4.57	
x2	-3.52984	1.048	-3.37	
x3	-2.20055	1.320	-1.67	

Note: Using equation 2, Tc, or the MAT estimate for San Jose = 21.3 °C

8/99 Kathryn Gregory-Wodzicki

[3] See footnote 1 for explanation of why 20 nearest neighbors are recommended.