**Example of an Ecological Data Set**

**Oak Woodlands in the Willamette Valley, Oregon, USA**

In 1961 and 1962 John F. Thilenius sampled vascular plants in oak forests in the Willamette Valley for his Ph.D. at Oregon State University (Thilenius 1963, 1968). The data came from a fairly narrow range of habitats – all of the stands were closed forests dominated by *Quercus garryana*. This resulted in a data set with fairly low beta diversity. The environmental differences among the sites are rather modest. Much of the variation in species composition presumably is derived from the particular histories of each stand, such as episodes of grazing, logging, and fire. Of course we have limited information on those histories, so you will see that much of the variation in the plant communities is not readily explained by the measured environmental and historical variables. Nevertheless a definite environmental gradient emerges from the analysis.

The abstract from Thilenius (1968) is reproduced below:

“*Quercus garryana* forests, prominent at low elevations throughout the Willamette Valley, Oregon, have developed from oak savanna subsequent to settlement of the valley in the mid-nineteenth century. Interruption of the ground fires that were common in the pre-settlement environment probably caused the change. The understory of the oak forest is dominated by shrubs, and well-defined strata are present. Four plant communities occur: (1) *Quercus garryana/Corylus cornuta* var. *californica/Polystichum munitum* (most mesic); (2) *Quercus garryana/Prunus avium/Symphoricarpos albus*; (3) *Quercus* *garryana/Amelanchier alnifolia*; (4) *Quercus garryana/Rhus diversiloba* (most xeric). All are in seral condition because of their relatively recent development and because they have been disturbed throughout their existence by man’s activities. The soils supporting the oak forest are generally deep and well drained and have developed profiles with illuvial horizons and acidic reaction. They are derived from sedimentary and basic igneous rocks and old valley-filling alluvium. Seven established soil series are present: Steiwer, Carlton, Peavine, Nekia, Dixonville, Olympic, and Amity. The Steiwer series and its catenary associate, Carlton, are the most common soils.”

Thilenius’ goals were to describe “the floristic composition, stand structure, physical environment, and successional status of plant communities where *Quercus garryana* is the major component of the overstory.” Although quantitative data were carefully recorded, Thilenius had few possibilities for multivariate analysis. His primary analyses were first arranging his data “according to similarities in species composition, importance ranks, and environmental attributes.” He then tabulated averages for species and environmental variables within the four groups.

Here is an interesting challenge for modern community analysts: what can you add to his account (Thilenius 1968) based on a more sophisticated quantitative analysis of the data? I mentioned above that a single strong environmental gradient emerges from the analysis, but this is only hinted in Thilenius’ abstract. What is that gradient?

After a listing of the files and variables contained in the files, three example procedures are given. The first demonstrates modification of the raw data into a form suitable for analysis. The second is an ordination with nonmetric multidimensional scaling. The third compares groups of sample units, as defined by landform.

**Files Provided**

**OakWoods.docx** – Microsoft Word file containing this document.

**OakWood1.mjm** (and same thing as \*.xlsx) – Main matrix containing species abundances in a matrix of 47 stands x 103 species. Abundances were derived from basal areas for trees and canopy cover for other species. Some species are listed more than once because abundance was evaluated separately for different height classes, as indicated by the suffixes: t = tree, s = shrub. Species with fewer than three occurrences were deleted, then the matrix was relativized by species maximum (i.e. each element in the matrix is expressed as a proportion of the maximum value in each column). Relativization by columns is necessary for some analyses because different columns were measured in different units. See OakRaw.xlsx for the raw data.

**OakWood2.mjm** (and same thing as \*.xlsx) – Second matrix of 47 stands x 27 attributes. The attributes are described in detail below. They include environmental variables, indicators of stand history. I have also added some community summary variables, including species richness, groups derived from cluster analysis, and community types as originally designated by Thilenius.

**OakRaw.xlsx** – The raw data matrix containing 47 stands x 189 species, before any modifications. The values are basal areas (ft2/acre) for trees and percentage cover for lower strata, based on 60, 0.2 m2 quadrats/stand. “Trace” was converted to 0.5%. A check on the field data sheet was converted to 0.2%. Be careful! Any use of these raw data must recognize that the columns representing the tree stratum differ in units from the lower strata; hence, use of a relativized matrix in OakWood1.xlsx.

**OakWoodsSimple.xlsx** – Same as OakWood1 and OakWood2, but in simple spreadsheet format (column names in first row), with each matrix a different worksheet ("Species" and "SecondMatrix")

More on the methods from Thilenius (1968):

“Investigations were confined to closed-canopy stands 4 ha or more in area where *Quercus garryana* was the major component of the overstory. Basal area, frequency, and density of overstory trees were determined on twenty 0.004-ha circular plots spaced at 9-m intervals in four rows parallel to the slope contour. Density was recorded in four classes: saplings (< 10 cm dbh); poles (11-40 cm dbh); mature (41-100 cm dbh) and relict (> 100 cm dbh). The maximum height of trees on each plot was measured with an optical rangefinder.”

“Frequency and percentage crown coverage of shrub and herbaceous species were recorded on sixty 0.2 m2 quadrats spaced at 3-m intervals in four rows coincident with the rows of 0.004-ha plots. Very low crown coverage was recorded as trace and arbitrarily assigned a value of 0.5% for calculation purposes. Above trace, the intervals were 1% and 5%. Coverage greater than 5% was estimated to the nearest 10%.”

**Coding for Variables in the Second Matrix**

**Topographic and geographic variables**

Elev = elevation above sea level in meters.

LatAppx = approximate latitude, decimal degrees, based on automated conversion of Township/Range/Section, using the program TRS2LL.exe.

LongAppx = approximate longitude, decimal degrees, based on automated conversion of Township/Range/Section, using the program TRS2LL.exe.

SlopeDeg = slope in degrees (originally recorded in percentages)

AspClass = aspect class, 1=SW, 2=S or W, 3=SE or NW, 4=N or E, 5=NE.

AspDeg = aspect in degrees E of N

PDIR = Potential annual direct incident radiation, MJ/cm2/yr, calculated according to McCune and Keon (2002) Eq. 3.

HeatLoad = Heat load index, calculated according to McCune and Keon (2002).

Landform: 1=valley bottom, 2=draw or slope of draw, 3=slope, 4=ridge

TopoClas = Topographic position class: adapted from scales used by Whittaker & Kessell (Kessell 1979):



**Soil variables**

Drainage: 1=poor, 2=moderate, 3=good, 4=well.

Soil series: 1=Steiwer, 2=Peavine, 3=Dixonville, 4=Nekia, 5=Carlton, 6=Olympia, 7=Amity

SoilGrp: 1=sedimentary, 2=basic igneous, 3=alluvial

A-horiz = thickness of A horizon, cm

B1-horiz = thickness of B1 horizon, cm

B2-horiz = thickness of B2 horizon, cm

B3-horiz = thickness of B3 horizon, cm (if profile truncated, e.g. “44+ inches”, add 20 inches)

B-horiz = sum of B1+B2+B3, cm

**Indicators of stand history**

GrazCurr = signs of current grazing recorded on field data sheet (0=no, 1=yes)

GrazPast = signs of past grazing recorded on field data sheet (0=no, 1=yes, must be 1 if GrazCurr=1)

NotLoggd = NPL recorded under “Influences” on data sheet. I guessed this means “no past logging”, i.e. no signs of past logging.

Que>60cm = number of *Quercus garryana* recorded in the 60 cm (24 inch) size class and larger. (No stands had large *Pseudotsuga*; one stand (Stand05) had a large *Acer macrophyllum* and one stand (Stand07) had two large *Arbutus menziesii*)

LogQ>60 = log of (*x*+1) where *x* is the number of *Quercus garryana* recorded in the 60 cm (24 inch) size class and larger (i.e. x = “LogQ>60”).

TreeHtM = maximum height of *Quercus garryana* in meters.

**Community summary variables derived from the species matrix**

SppRich = species richness, calculated from OakRaw.xlsx, counting each species x layer combination as a separate species.

ThilType = vegetation types from Thilenius (1968)

1 = *Quercus/Corylus/Polystichum*

2 = *Quercus/Prunus/Symphoricarpos*

3 = *Quercus/Amelanchier/Symphoricarpos*

4 = *Quercus/Rhus*

FlxB-.25 = community types defined at the 4-group level from hierarchical cluster analysis, Flexible beta method, Sørensen distance, beta= -0.25.

**List of Species Codes**

Note: because woody species may occur in more than one stratum, a suffix (-s, -t) is used to indicate a given species in the shrub or tree stratum.

Abgr‑s Abies grandis SHRUB

Abgr-t Abies grandis

Acar Actea arguta

Acgld Acer glabrum var. douglasii

Acma‑s Acer macrophyllum shrub

Acma-t Acer macrophyllum

Acmi Achillea millefolium

Adbi Adenocaulon bicolor

Agha Agrostis hallii

Agre Agropyron repens

AGRO Agrostis sp?

Agse Agrostis semiverticullata (subsecundum)

Agte Agrostis tenuis

Aica Aira caryophyllea

ALL Allium sp.

Alpr Alopecurus pratensis

Amal‑s Amelanchier alnifolia shrub

Amal-t Amelanchier alnifolia

Apan Apocynum androsaemifolium

Aqfo Aquilegia formosa

Arel Arrhenatherum elatius

Arme‑s Arbutus menziesii SHRUB

Arme-t Arbutus menziesii

Avfa Avena fatua

Beaq Berberis aquifolium

Brpu Brodiaea pulchella

Brco Bromus commutatus

Brla Bromus laevipes

Brri Bromus rigidus

Brse Bromus secalinus

Brst Bromus sterilis

Brvu Bromus vulgeris

Caqu Camassia quamash

CAR Carex sp.

Cato Calochortus tolmiei

Cear Cerastium arenses

Ceum Centaurium umbellatum

Ceve Ceanothus velutinus

Cipa Circaea pacifica

Civu Cirsium vulgare

Coco‑s Corylus cornuta shrub

Coco-t Corylus cornuta

Cogr Collomia grandiflora

Conu‑S Cornus nuttallii SHRUB

Conu-t Cornus nuttallii

CORY Corylus sp.

Cost Corallorhiza striata

Crca Crepis capillaris

Crdo‑t Crataegus douglasii

Crdo‑s Crataegus douglasii

Crox Crataegus oxyacantha

Cyec Cynosurus echinatus

Cyfo Cystopteris fragilis

Cygr Cynoglossum grande

Daca Danthonia californica

Dacar Daucus carota

Dagl Dactylus glomerata

Deel Deschampsia elongata

Diar Dianthus armeria

Doel Downingia elegans

Drar Drysopterus arguta

Drar Dryopteris arguta

Elgl Elymus glaucus

Erla Eriophyllum lanatum

Erog Erythronium oregonum

Eucr Euphorbia crenulata

Feca Festuca californica

Fede Festuca dertonenses

Feel Festuca elatior var. arendmaceae

Feme Festuca megalura

Feoc Festuca occidentalis

Feru Festuca rubra

Frbr Fragaria bracteata (vesca)

Frcu Fragaria cuneifolia

Frla‑s Fraxinus latifolia shrub

Frla-t Fraxinus latifolia

Frvi Fragaria virginiana

GAL Galium sp.

Gema Geum macrophyllum

Geog Geranium oreganum (incisum)

Gepu Geranium pusillum

Haob Habenaria orbiculata

Haun Habenaria unalacensis

Hehe Hedera helix

Hemi Heuchera micrantha

Hodi Holodiscus discolor

Hola Holcus lanatus

Hyoc Hydrophyllum occidentale

Hype Hypericum perforatum

Hyra Hypochaeris radicata

Irte Iris tenax

JUNC Juncus sp.

Kocr Koeleria cristata

Laco Lapsana comunis

Lapo Lathyrus polyphyllus

Lasa Lathyrus sativus (Pisum sativum)

Liap Ligusticum apiifolium

Libu Lithophragma bulbifera

Lico Lilium columbianum

Lide-t Libocedrus deccurens

Lide‑s Libocedrus deccurens

LILI Lilium sp.

Loci Lonicera ciliosa

Lope Lolium perenne

LOT Lotus sp.

Lotr Lomatium triternatum

Lumu Luzula multiflora

Maex Madia exigua

MAL Malvaceae sp.

Maor Marah oreganus

Mebu Melica bulbosa

Mila Microseris laciniata

Mope Montia perfoliata

Mosi Montia sibirica

Nepa Nemophylla parviflora

ONGR Onagraceae sp.

Osce‑t Osmaronia cerasiformis tree

Osce-s Osmaronia cerasiformis

Osch Osmorhiza chilensis

Osnu Osmorhiza nuda (chilensis)

Phca Physocarpus capitatus

Phle Philadelphus lewisii

Phpr Phleum pratense

Phvi Phoradendron villosum

Pipo‑s Pinus ponderosa

Pipo Pinus ponderosa

Plla Plantago lanceolata

Poco Poa compressa

Pogl Potentilla glandulosa

Pogr Potentilla gracilus

Pogr Potentilla gracilis

Pomu Polystichum munitum

Popr Poa pratensis

Povu Polypodium vulgare

Prav‑s Prunus avium shrub

Prav-t Prunus avium

Prde-t Prunus virginiana var. demissa

Prde‑s Prunus virginiana var. demissa shrub

Prvu Prunella vulgeris

Psme‑s Pseudotsuga menziesii shrub

Psme-t Pseudotsuga menziesii

Ptan Pterospora andromedia

Ptaq Pteridium aquilinum var. lanuginosum

Pyco‑s Pyrus communis shrub

Pyco Pyrus communis

Pyfu‑s Pyrus fusca SHRUB

Pyfu-t Pyrus fusca

Quga-s Quercus garryana shrub

Quga Quercus garryana

Raoc Ranunculus occidentalis

Rhdi Rhus diversiloba

Rhpu‑s Rhamnus purshiana shrub

Rhpu Rhamnus purshiana

Risa Ribes sanguinius

Rodu Rosa???

Roeg Rosa eglanteria

Rogy Rosa gymnocarpa

Ronu Rosa nutkana

Ropi Rosa pisocarpa

Ropi Rosa pisocarpa

Ruac Rumex acetosella

Rula Rubus laciniatus

Rule Rubus leucodermus

Rupa Rubus parvifloris

Rupr Rubus procerus

Ruur Rubus ursinus

S‑2 Carex sp2.

S‑1 Carex sp1.

Sacr Sanicula crassicaulis

Sado Satureja douglasii

Sagr Sanicula graveolens

Seja Senecio jacobaea

Siho Silene hookeri

Smra Smilacina racemosa

Smse Smilacina sessilifolia

Syal Symphoricarpus albus

Taas Taeniatherum asperum

Taof Taraxacum officinale

Tegr Tellima grandiflora

Thoc Thalictrum occidentale

Toar Torilis arvensis

Trca Trisetum canescens

TRIF Trifolium sp

Trla Trientalis latifolia

Trov Trillium ovatum

Trpr Trifolium procumbens

V#1 Vicia sp.

Valo Valerianella locusta

Viam Vicia americana

Viel Viburnum ellipticum

Vinu Viola nuttallii

VIOL Viola sp

Zice Zygadenus venosus

**EXAMPLE ANALYSES**

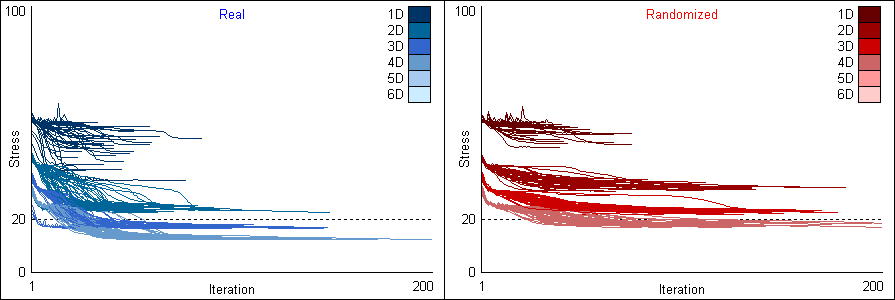
**Derivation of an adjusted data matrix (OakWood1) from the raw data matrix (OakRaw).**

(Note: for more on the rationale behind these steps, see McCune & Grace (2002), “Analysis of Ecological Communities.”

1. *File | New Project* | then enter a project name such as Oak Woods, select the folder with your data as User Data Folder
2. Drag the file OakRaw.xlsx into the main matrix position, then click *Save Project File*. (Other checkboxes can be left at the defaults)
3. Select Excel Spreadsheet as the Type of matrix
4. PC-ORD will automatically want you to save your main matrix as OakRaw.mjm, so be sure that your User Data Folder is selected and click "OK".
5. Delete species with fewer than three occurrences (*Modify Data | Delete Columns | Fewer than N Non-zero Values |* select *N=3* (The rationale for this is explained by McCune & Grace 2002, pp. 75-76)
6. Click *OK* in answer to *Do you wish to use Temp.mjm as the new Main Matrix?*
7. Note that the result file now shows a list of the 86 columns (species) that were deleted.
8. *Modify Data | Relativizations | Relativization by Maximum |* select *Columns: Species | OK.* (It is essential to relativize by columns (species) in this case, because some species have abundances as basal areas and some as percent cover; see p. 73 in McCune & Grace 2002).
9. You will be asked, *Current temporary RESULT.TXT file will be lost. Save file now?* Click on *Discard* (you do not normally need to save this, as it just has the list of the species that were deleted.)
10. Click *OK* in answer to *Do you wish to use Temp.mjm as the new Main Matrix?*
11. You will be asked, *Current temporary work file WORK.MJM will be lost. Save file now?* Click on *Discard* – no need to save this file. It contains the matrix after the infrequent species were deleted but before the relativization.
12. The main matrix should now be relativized by species maximum and contain 47 rows and 103 columns. The contents should be identical to OakWood1.xlsx.

**Nonmetric multidimensional scaling of the community data, with overlays from the second matrix.**

1. Open the file OakWood1.mjm as the main matrix (*File | Open | Main matrix*).
2. Select *Ordination | NMS.*
3. Select the Autopilot tab, check Autopilot mode, and select *Medium.* (If you have a fast computer you might wish to select *Slow and Thorough.*
4. On the Distance Measure tab, select *Sørensen*. (The selections on the other tabs cannot be set because you have turned on autopilot. Any options previously selected on those tabs will be ignored.)
5. Click *OK.*
6. Enter a descriptive title for the results, for example, “Thilenius data, NMS medium thoroughness,” then click *OK*.
7. If unsaved results from a previous action are showing in the result window, you will be asked, *Current temporary RESULT.TXT file will be lost. Save file now?* Click on *Discard* or one of the other options, depending on what you want.
8. NMS will run, using 50 random starts with the real data set and 50 random starts using different randomizations of the data (shuffling within columns). For each starting configuration NMS will seek a stable 4-, 3-, 2-, and 1-dimensional solution. Real-time graphs of stress for each run are displayed, while your computer is working on the problem.



1. When the run is complete, close the stress graph, then a new result file will appear, along with windows containing coordinates for the stands (GRAPHROW.GPH) and the species (GRAPHCOL.GPH). Save each of these under a new name using *File | Save as | All,* then enter a Base Name, for example "NMSThil". This name will then be used as a basis for naming your result file and the row and column coordinates, for example as NMSThilResult.txt, NMSThilRow.gph and NMSThilSppCol.gph.
2. Inspect the result file. See the chapter in McCune & Grace (2002) on NMS. Because random starts are used, your results will differ somewhat from those given here. A key portion of the results file is the following table.

STRESS IN RELATION TO DIMENSIONALITY (Number of Axes)

--------------------------------------------------------------------

Stress in real data Stress in randomized data

50 run(s) Monte Carlo test, 50 runs

------------------------- -----------------------------------

Axes Minimum Mean Maximum Minimum Mean Maximum p

--------------------------------------------------------------------

1 34.486 49.777 56.481 47.752 54.478 56.485 0.0196

2 22.609 23.664 25.109 29.697 31.914 34.104 0.0196

3 16.419 16.771 17.741 21.380 22.776 24.139 0.0196

4 12.320 12.396 12.918 16.923 17.810 19.277 0.0196

--------------------------------------------------------------------

p = proportion of randomized runs with stress < or = observed stress

i.e., p = (1 + no. permutations <= observed)/(1 + no. permutations)

Conclusion: a 3-dimensional solution is recommended.

Now rerunning the best ordination with that dimensionality.

1. Note that the *p*-values indicate that solutions of any dimensionality from 1 through 6 are stronger than expected by chance. Autopilot chose a 3-D solution because it reduces the stress by over 5 units, versus a 2-D solution, while the giving a small *p*-value. The final stress for the best 3-D solution was 16.4.
2. Open the second matrix so you can study the relationship between those variables and the community structure: *File | Open | Second Matrix | OakWood2.mjm*.
3. View the ordination graph. Select *Graph | Graph Ordination* See Chapters 13 and 16 in McCune & Grace (2002) for suggestions on how to interpret the results. For example you might wish to:
   1. display a joint plot (*Graph | Joint Plot*), and examine each pair of axes,
   2. see how much of the variation in the distance matrix is represented in the ordination diagram (*Statistics | Percent of Variance in Distance Matrix*),
   3. graphically examine the relationships between the ordination and individual variables in the second matrix (*Graph | Overlay From Second Matrix*),
   4. calculate linear and rank correlation coefficients between axis scores and variables in the second matrix (*Statistics | Correlations With Second Matrix*),
   5. rotate the diagram so that major vectors in the joint plot are aligned with the axes (select *Graph | Joint* plot, then *Rotate | By Angle Continuous*. Select 5 degrees for the increment and click *Next* repeatedly to gradually rotate the ordination. See Chapter 15 in McCune and Grace (2002). If you wish to save your rotation, select *File | Save Scores As | Rows*: *Stands | Text File*, then choose a filename, such as NMSThilRot.gph.
   6. explore the options – there is a lot here. Take the time to familiarize yourself with the various menu items and options.

**Comparison of communities among groups of stands as defined by categorical variables.**

1. Open the file OakWood1.mjm as the main matrix (*File | Open | Main matrix*).
2. Open the second matrix because it contains the grouping variables: *File | Open | Second Matrix | OakWood2.mjm*.
3. Select *Groups | MRPP | Define Classes From Second Matrix*.
4. Select one of the categorical variables, for example “Landform” as the grouping variable.
5. In the MRPP Setup dialog, select the *Sørensen* distance measure, n/sum(n) as the weighting method, and check the box, *Rank transform distance matrix*.
6. Click *OK*.
7. Enter a descriptive title for the results, for example, “Thilenius data, rank-transformed MRPP, Sorensen distance”, then click *OK*.
8. If unsaved results from a previous action are showing in the results window, you will be asked, *Current temporary RESULT.TXT file will be lost. Save file now?* Click on *Discard* or one of the other options, depending on what you want.
9. MRPP then runs, writing the results into RESULT.TXT.
10. Inspecting the results window, you should see the following.

Chance-corrected within-group agreement, A = 0.07630772

Probability of a smaller or equal delta, p = 0.00772772

This *A* value tells us that species composition differs only modestly among landforms, while the *p*-value tells us that it is quite unlikely for us to obtain this result just by chance. For more help interpreting the results from MRPP, see Chapter 24 in McCune and Grace (2002).

1. If you wish to see which species underlie the differences among landforms, try Indicator Species Analysis. From the main menu, select *Groups | Indicator Species Analysis | Define Classes From Second Matrix*.
2. Again, select Landform as the grouping variable.
3. In the Setup dialog, select Quantitative data, uncheck *Exclude one or more groups* and check *Randomization test*.
4. For the random numbers, select *Use time of day*. Click *OK*.
5. Select *Number of runs = 1000*, then *OK*.
6. Enter a descriptive title for your results, to be used as a header in the results file, then click *OK*.
7. If unsaved results from a previous action are showing in the results window, you will be asked, *Current temporary RESULT.TXT file will be lost. Save file now?* Click on *Discard.*
8. The results from ISA will appear in the results window. See Chapter 25 in McCune & Grace (2002) for help in interpreting the results.
9. One particularly good indicator of landform is “Arel” (*Arrhenatherum elatius*), which has the following result in the final table:

---------------------------------------------------

IV from

Observed randomized

Indicator groups

Column Maxgrp Value (IV) Mean S.Dev p \*

------------- -------- -------- ----- ------ -------

15 Arel 1 86.5 16.6 11.47 0.0050

The Maxgrp=1 tells us that it is particularly consistent in Landform 1 (valley bottom). Looking back at the preceding tables, we see that it has a high concentration of abundance in group 1 (86% in the Relative Abundance table) and it was always present in that group (100% in the Relative Frequency table). You p-value may not be exactly the same as this, because you probably had a different sequence of random numbers for the monte carlo test.

**Acknowledgments**

The original raw data cards from Thilenius' study in 1963 (data collected in 1961 and 1962) were obtained from John Thilenius via Bob Frenkel. Thanks to John Thilenius for granting permission to distribute his data. Bill Daly did the initial data entry. Bibit Traut added more variables and resolved numerous nomenclatural questions regarding the species codes used by Thilenius.

**References**

Kessell, S. R. 1979. Gradient Modeling: Resource and Fire Management. Springer-Verlag, New York. 432 pp.

McCune, B. and D. Keon. 2002. Equations for potential annual direct incident radiation and heat load. Journal of Vegetation Science 13:603-606,

Thilenius, J. F. 1963. Synecology of the white-oak (*Quercus garryana* Douglas) woodlands of the Willamette Valley, Oregon. Ph.D. Dissertation. Oregon State University, Department of Botany and Plant Pathology, Corvallis. 151 pages.

Thilenius, J. F. 1968. The *Quercus garryana* forests of the Willamette Valley, Oregon. Ecology 49:1124-1133.