Data Analysis with Python
Final project assignment
Last update: 20.12.2019
Created by: Jaakko Toivonen, Indre Zliobaite
Contact: jaakko.toivonen@helsinki.fi

1 Detecting patterns of speciation in the fossil record

In this assignment, we use data from the NOW (New and Old Worlds) database of fossil mammals to study patterns of speciation over time and space. In particular, we are interested to know when and where speciation rates have been significantly high. The task is to find which time periods and which places over the history of mammals have given rise to exceptionally high numbers of new species. The phenomenon is known in the evolutionary literature as the "species factory". Palaeontologists are interested why and in which ways those times and places are special. The role of computational science is to identify and characterize such times and places.

We practice using pandas DataFrames, performing logistic regression and making statistical significance tests in data analysis.

1.1 Fossil data and the NOW database

Fossils are remains, traces or impressions of organisms that lived in the past, preserved in rocks. The NOW fossil mammal database contains global information about Cenozoic land mammal taxa (identifications of animals and their ancestral relationships to other animals) and localities (places on Earth where fossils are found). Cenozoic is the era that extends from 66 million years ago to the present day, that is, a time interval starting right after dinosaurs went extinct. Cenozoic is known as the Age of Mammals, because the extinction of many groups, including dinosaurs, allowed mammals to greatly diversify. The continents also moved into their current positions during this era.

The NOW database has been curated in Helsinki since 1993. The database keeps a record of global mammal fossil finds: it records geographic localities where mammalian fossils have been found around the world and for each locality there is a list of species that have been found there. The database

includes an estimated age for each fossil locality. Information about fossil mammal species, characteristics of their appearance, behaviour and ways of life is also available. The data is free and open to use for anyone.

The database has been compiled over many years from information given in publications and from personal knowledge of domain experts. The database is alive, meaning that data is continuously updated and expanded to incorporate new discoveries. The database is curated by an international team of experts. However, fossil data is inherently uncertain. Taxonomic assignments as well as age estimates rely on opinions and interpretations of human experts.

There are different ways to estimate the age of a fossil. Typically, the age is assigned based on the age of the locality at which the fossil was found. One method to do this is through radiometric dating. This relies on the speed of radioactive decay of chemical elements and can tell age very precisely. The main challenge is that the sample must have remained a closed system, for instance, due to a volcanic eruption, since the event being dated. Some continents and some time periods offer more possibilities for radiometric dating than others. The African fossil record, for instance, is quite well interleaved with volcanic layers, while the European fossil record – not so much. In the absence of other possibilities, relative dating is used to get age estimates. This technique relies on common biological events, such as first and last occurrences of selected indicator species. Such dating is called biochronology: the age of a locality is estimated by comparing the faunal composition of the locality to known reference localities, for which exact ages are known via radiometric dating. Mammal Neogene (MN) time zones in Europe is a set of such relative dating units. We will use those units in this analysis.

Global compilations of fossil data can be used for different types of analyses of the history of life, evolutionary processes and environmental contexts. The goal is not only to repicture ecosystems of the past from a highly fragmented record, but more so to reconstruct processes and drivers of changes in those ecosystems. The fundamental question that can be asked of such data is how life on Earth works.

1.2 Data preprocessing

This exercise will focus on analysis of how new species appear and how those patterns manifest in space and time. We will work with time intervals corresponding to the European MN zones. As part of the data preprocessing

exercise we will need to map Asian and African fossil localities to these same units, so that we are working with one unified time frame.

The next steps will guide you through preprocessing data before proceeding to computational analysis.

Exercise 1. Download data from the NOW database. Go to

http://www.helsinki.fi/science/now/

Click "View Database", then "Enter Database", then "Locality" and then "Export". Select "include species lists" and choose "Comma" for field separator. Then, click "All NOW localities". Once download completes, copy all of the text on your browser and paste it into a new txt file. Save the file.

Once we have all of the raw data, we need to transform it into a more useful format.

Exercise 2. Create a pandas DataFrame that contains all of the data and save it as a csv file. How many rows does the DataFrame contain?

In the DataFrame, each row represents one fossil occurrence. The columns LONG and LAT give the longitude and latitude, respectively, of where a given fossil occurrence was found. The columns MIN_AGE and MAX_AGE give an estimation of the age of the fossil (in millions of years). The column LIDNUM contains a unique identification number for each locality, where fossils have been found. The columns GENUS and SPECIES give information on the taxonomic identification of the fossil.

The modern modular system of naming organisms was invented by Carl Linnaeus in the 18th century. This system is exclusively used today. Within the scope of this exercise, it is important to know that a species is uniquely identified by a combination of the above two taxonomic ranks: genus and species. Neither rank alone uniquely defines a species, but their combination is sufficient. As a side note, a species name should always be written in *italics*, for example *Canis lupus* (the gray wolf), where *Canis* is the genus and *lupus* is the species.

Before we can begin to analyze the data properly, a number of cleaning and preprocessing steps are necessary.

Exercise 3. a) Remove all rows where LAT = LONG = 0; these occurrences have incorrect coordinates. Drop rows where SPECIES is "sp." or "indet."; these occurrences have not been properly identified.

name	max	min
MN1	23	21.7
MN2	21.7	19.5
MN3	19.5	17.2
MN4	17.2	16.4
MN5	16.4	14.2
MN6	14.2	12.85
MN7-8	12.85	11.2
MN9	11.2	9.9
MN10	9.9	8.9
MN11	8.9	7.6
MN12	7.6	7.1
MN13	7.1	5.3
MN14	5.3	5
MN15	5	3.55
MN16	3.55	2.5
MN17	2.5	1.9
MQ18	1.9	0.85
MQ19	0.85	0.01

Table 1: MN time unit boundaries (in millions of years ago). Note that MN7 and MN8 are combined into one time unit. This is a decision that scientists have made because of complications in separating the two units at popular fossil localities.

b) Next we will assign each occurrence to a specific Mammal Neogene (MN) time unit. Table 1 shows the time boundaries of each time unit. Assign each occurrence to a correct time unit by calculating the mean of MIN_AGE and MAX_AGE. If the mean age of an occurrence is precisely on the boundary between two time units, assign the occurrence to the older time unit. If the mean age of an occurrence is outside of the MN time interval, assign it to a "pre-MN" or "post-MN" category.

c) Sometimes expert knowledge may be used to override some of the information recorded in the data. In our case, experts in palaeontology tell us that occurrences in the localities "Samos Main Bone Beds" and "Can Llobateres I" should be assigned to time units MN12 and MN9, respectively. Check these and if necessary, edit the time units to their correct values.

d) We need to be able to identify all occurrences of each species. Assign

a unique identification number for each unique combination of GENUS and SPECIES. Create a new column in the DataFrame and label each occurrence with a corresponding species identification number.

e) Each locality should contain no more than one occurrence of any species. Check whether this is the case and remove duplicate copies, if necessary.

f) How many rows are we left with in the DataFrame (compare with exercise 2)? How many unique species and localities are identified?

1.3 Occurrences

For each species in our dataset, we can determine the time unit when that species is first observed in the fossil record within our study area. We refer to all of the occurrences of a given species in the oldest time unit it is observed in as first occurrences. Then, we can translate our interest in high speciation rates into detecting when and where we observe a lot of first occurrences.

Exercise 4. Create a DataFrame that shows for each species how many occurrences it has in each time unit. Then, create a different DataFrame that shows for each species the time unit when it is first observed (i.e. the oldest time unit). For each time unit, calculate the proportion of first occurrences to all occurrences. Plot the proportion of first occurrences over time. Also, plot the total number of occurrences over time.

Next, we would like to look into geographic patterns in the data. A useful library for plotting geographic data is geopandas (http://geopandas.org/). To import a world map, use the following code:

```
import geopandas
world = (geopandas.read_file(geopandas.datasets.get_path
        ('naturalearth_lowres')))
```

For plotting, you may use, for example, the following code:

```
fig, ax = plt.subplots(figsize=(20,10))
world.plot(ax=ax, color='wheat', edgecolor='black')
ax.set_facecolor('azure')
```

Exercise 5. a) Create a DataFrame that collects the following information for every locality: locality number (LIDNUM), longitude, latitude, time unit, number of first occurrences in the locality, number of all occurrences in the locality and proportion of first occurrences in the locality.

Note, you should use LIDNUM to identify unique localities and not the NAME variable (why?).

b) Visualize the distribution of localities in space and time. For each time unit, plot the LAT and LONG coordinates of each locality (corresponding to the time unit). For example, you can use the above codes to create a geographic map and then use a standard matplotlib scatter plot to add the localities. Choose the marker size for each locality such that it is relative to the number of occurrences in the locality (bigger markers for bigger localities).

c) Based on exercises 4 and 5, what kind of observations about sampling can you make? Are there differences in sampling density over space and time? Compare some basic sampling properties between Africa, Asia and Europe, e.g. spatial coverage and average number of occurrences per locality.

1.4 Localities and sampling

One possible way to reduce noise in the fossil data is to spatially aggregate occurrences. By this we mean the following: instead of considering only the occurrences of a given locality, we consider all occurrences found in some geographic area around the focal locality (encompassing other fossil localities) to be representative of the fauna that lived in the area. After all, given the typical length of MN time units (around 1 million years) and the mobility of large mammals, it is reasonable to assume that occurrences in nearby localities should also be reflected in our estimation of the faunal composition of the focal locality.

The number of first occurrences observed in a given time and place is dependant on how well that area was sampled in previous time units. For example, we would expect to find more first occurrences in an area that was poorly sampled in the preceding time unit, because it is more likely that we simply do not have observations of some species in the previous time unit that nevertheless existed then.

A general challenge with fossil data is that it suffers from a sampling bias, i.e. sampling is uneven in space and time. Then, it is not very informative to simply look when and where you find most first occurrences, because the number of first occurrences is correlated with the total number of occurrences in a given time and area (which reflects sampling bias).

Next, we will measure sampling density for each locality as the number of occurrences observed in a given area around that locality in the preceding time unit. We will also collect occurrence statistics for that same area in the time unit of the locality. These statistics will later be used to estimate statistical significance of the observed number of first occurrences.

Exercise 6. For each locality, look at a ten by ten degrees area (in latitude and longitude) centered around the locality. Record the total number of occurrences and total number of first occurrences found within that square in the time unit corresponding to the focal locality. Also, record the total number of occurrences within that square in the preceding time unit (relative to the focal locality). Record these numbers into the DataFrame that was created in exercise 5 (add new columns).

1.5 Logistic regression

In order to determine if an observed number of first occurrences in a given area at a given time is significantly high, we should first establish a reasonable expectation for that number. Then, we can compare our observations to the expectation to find out when and where there are significantly many first occurrences observed. So, we want to look at how the number of observations (occurrences) in a given focal area in the preceding time unit affects the proportion of first occurrences found in the focal locality now. To this end, we will perform a simple regression analysis. In particular, given that the data here is in the form of "first occurrence" or "not a first occurrence", we will use logistic regression. In this manner, we will be able to establish what is a reasonable expectation for the proportion of first occurrences observed now given past sampling density.

For the logistic regression, we use the statsmodel library (https://www.statsmodels.org/stable/index.html). The following is an example code for how to use statsmodels:

```
import statsmodels.api as sm

# X gives the data for the independent variable
X = regression[:,0]

# add a constant term to the regression (see below)
X = sm.add_constant(X)

# y gives the data for the dependent variable
y = regression[:,1]
```

```
# logistic regression
model = sm.Logit(y, X)
result = model.fit()
print(result.summary())

# get the estimated parameters for the logistic regression
coefficients = result.params

# get the estimated parameters for the 95%-confidence interval
# of the logistic regression
confidence95 = result.conf_int(alpha=0.05)
```

In the above code, we fit a logistic regression curve of the form

$$y(x) = \frac{1}{1 + e^{-(c_0 + c_1 x)}}, \tag{1}$$

where x is the independent variable (past sampling density), y the dependent variable (expected proportion of first occurrences now) and c_0 and c_1 are the parameters solved by logistic regression (sm.add_constant() adds the c_0 term to the regression).

From here on we will be focusing on an area roughly covering Europe, which is relatively well sampled over time.

Exercise 7. a) Create the regression data set. Only use localities within the co-ordinates -25<LONG<40 and LAT>35 and time unit within MN2-MQ19 (why not include MN1?). Create an $m \times 2$ array, where m is the total number of occurrences in all the localities. Each row in the array represents one occurrence. For each occurrence, fill in to the first column of the array the number of occurrences in the focal area in the previous time unit (calculated in exercise 6). For the second column, fill in 1 for a first occurrence and 0 for other occurrences.

b) Perform logistic regression.

c) Plot regression curve and 95%-confidence intervals.

1.6 Statistical significance

Now that we have performed the logistic regression, we have a way to systematically evaluate how many first occurrences to expect, given sampling density in the previous time unit.

Exercise 8. For each European locality, calculate the expected proportion of first occurrences in the focal area surrounding the locality using the logistic regression calculated in exercise 7.

Now we know for each locality how many first occurrences there are in the area around the locality and we have an estimation of how many first occurrences to expect to find there. Then, we can perform a standard statistical test to evaluate whether the observed number of first occurrences is significantly higher than what we would expect (based on the regression).

Exercise 9. For each European locality, calculate the probability of observing as many or more first occurrences in the focal area than what is actually found. Assume that occurrences are binomially distributed to "first occurrences" and "not first occurrences", so that the probability of a given occurrence to be a first occurrence is equal to the expected proportion of first occurrences in the focal area. You may use, for example, the scipy.stats.binom library (https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.binom.html) for the calculations.

The above calculations give us a method to evaluate systematically, when and where the data shows an unexpectedly high amount of first occurrences, while taking into account variable sampling in space and time. Localities that show the probability of the observations to be 0.05 or less are typically thought of as statistically significant. We can look for interesting speciation patterns in space and time by plotting the localities with their significance indicated.

Exercise 10. For each time unit, plot localities on a map covering the coordinates defined in exercise 7a and indicate their significance level with a sliding color scheme. Highlight localities that have p-value less than 0.05 (i.e. probability of observations is less than 0.05). Describe briefly the overall patterns that you observe.