Assignment_1_Extracting_the_dominant_periods_of_a_time_series

February 14, 2019

0.0.1 This assignment consists of five parts:

1. Create synthetic data to test the algorithm you desing
2. Write two functions: (i) autocorrelation and (ii) period extraction function
3. Use a periodogram function (possibility for bonus points)
4. Compare sensitivity of the algorithms to typical imperfections that happen in real data (noise, missing data, random and non-periodic patterns,)
5. Test the algorithms on geolife data

Part 1. to 3. will make up 70 % of your grade, part 4. and 5. constitute 30% of your grade. There is a bonus that you can gain in part 3 (you'll get up to 10%).

WARNING: Make sure to read through the entire assignment before starting to code. The tasks build on each other!

Good luck with the assignment! Deadline is February 28th, at 23:55 hrs. Please hand in your .ipynb and .pdf via blackboard.

1st part: Create Synthetic Data

Imagine that you have GPS device set to take measurements every few time intervals. Simulate first clean trajectory data with two periodicities for 365 days. The trajectory data should include two types of periodicities (e.g. one weekly and one daily). For example you could take your Home-Snellius trajectory as a daily trajectory and your weekly trip to the supermarktet as the second one. Simulate the whole trajectory, path and the time you spend in each palce. You can find GPS locations via google maps (right click on map - what is here?) or any other online map service. Try to think of scenarios that make this data as accurate as possible. (You can be creative about your home location because that's private information!)

The GPS location of the entrance of the Snellius building is: 52.169709, 4.457111.

Use a constant time between way points/ GPS locations for this exercise.

Tip: a periodicity of 24 is recurring event at every 24 hrs (i.e. a daily event), a periodicity of 168 hrs is a weekly event. A time series that is 24-168 is data with a daily recurring event and one weekly recurring event. Tip: We talked so far about processing one time-series. In case two time-series acquired from two coordinates is difficult to handle, try to use only one, or combine them into one value (e.g. lat^2 + long^2)

```
In [0]: #simulate data:
```

```
In [0]:
```

0.0.2 2nd part: Write two functions

Your task is to write a) a function that performs an Autocorrelation on the simulated data and plots a Autocorrelation graph b) a function that evaluates the output of autocorrelation and extracts the two simulated periodicities as output. Construct the second function such that it will return two numbers as periodic components (24, 168) and indicate which periodicity is more prominent.

```
In [0]: #write your own ACF

In [0]: #display ACF on y-axis, time on x-axis

In [0]: #write your evaluation function
```

0.0.3 3rd part: Periodograms

Take a periodogram function from available functions in python and run it on the simulated data. Does your evaluation function from the 2nd part works on the results of the periodogram? If not desing another algorithm for extracting periods from the periodogram. Bonuspoints for writing your own periodogram function.

```
In [0]: #work with periodogram

In [0]: #evaluate results of periodogram
```

0.0.4 4th part: Performance

Noise can occur at different levels in this data:

- missing measurements at different proportions (randomly or in bursts)
- noise around the location data
- noise around the temporal data
- irregular behavior (i.e. skipping or adding a trajectory; for example going to school on a saturday or not going grocery shopping for a week)

Choose at least two noise sources, add them to your simulated data and compare the performance of your ACF and periodogram function. Parametarazie your process of injecting these uncertainties and check how sentitive your algorithms are to different proportions of these sources of imperfections. Which one performs best? Under which circumstances?

```
In [0]: #compare performance
```

Write here a little report on your findings (150 words max):

5th part: Real life data

You will use data from the geolife data sets to test the analysis methods that you worked with. Your task is to run the analysis methods you just learnt and interpret your findings.

Download the data here: https://www.microsoft.com/en-us/download/details.aspx?id=52367. Select one participant (the participants differ a lot in terms of number of trajectories collected, you might want to check several participants) and find out as much as you can about that participant.

The user guide of the entire dataset can be found here: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/User20Guide-1.2.pdf

Use these questions as a guideline for your interpretation:

- Where was the data gathered? Can you find frequently visited locations by exploring the data?
- What is the general structure of your data? How much noise do you expect? What is the temporal granularity of your data? How long did your participant log their movement for?
- Can you find periodic behaviors? Of which time?
- If you cannot identify periodic behaviors: can you say why? What makes your data challenging? What realistic aspects of data was missing in your simulation? Having these challenges in mind: If you were to design a data collection protocol, what would be your topmost priorities?

```
In [0]: #import data

In [0]: #some general exploratory statistics to find out more about your participar

In [0]: #run (your own) ACF

In [0]: #run periodogram on the data
```

Conclusion about the data, little report on findings (max 250 words):