Final exam

J.-S. Bourdeau and N. Saunier

December 20, 2017

Please

- note the scale (the total score is out of 20) and the indicative time to devote to each exercise;
- clearly indicate the numbers of the questions you are dealing with and your corresponding answers (and underline or frame the numerical results);
- pay particular attention to the wording and definition of the notations you use;
- download the data necessary for certain exercises available in the data archive on moodle (Section " Final exam ", first element " Data ") (the text files are provided in a version with a point and a comma for decimals, if necessary).

Statistical tables are available on moodle if necessary.

Exercise 1 (statistical analysis and regression model) 65 min (/ 9 pts)
This exercise is based on a data set from the City of Washington, D.C. bike-sharing system for the years 2011 and 2012 (file `washington-day.csv`). The attributes of the file are as follows:

- instant: identifier
- dteday: date
- season: season (1: spring, 2: summer, 3: fall, 4: winter)
- yr: year (0: 2011, 1: 2012)
- mnth: month (1 to 12)
- holiday: holiday indicator
- weekday: day of the week
- workingday: weekday indicator
- weathersit: weather conditions
  - 1: clear, few clouds
  - 2: cloudy, fog
  - 3: snow or light rain
  - 4: heavy rain, storm, heavy snow

- temp: normalized temperature in degrees Celsius (divided by 41)
- atemp: felt temperature in degrees Celsius, normalized (divided by 50)
- hum: normalized humidity (divided by 100)
- windspeed: normalized wind force (divided by 67)
- casual: number of bikes borrowed per day by occasional users
- registered: number of bikes borrowed per day by subscribers
- cnt: number of bikes borrowed per day (sum of " casual " and " registered ")

Please answer the following questions:

1. Propose a graph to study the correlation between the variables " casual " and " registered " according to the days of the week and the weekend. Comment on the graph and the correlations. (1.5 Pts)
2. Determine using a statistical test whether the number of bicycles borrowed per day (variable " cnt ") is greater on clear days than on cloudy days. (1.5 Pts)
3. What statistical test can be used to study the correlation between the number of bicycles borrowed per day (variable " cnt ") and weather conditions (variable " weathersit ")? What are the conditions of application of the test? (1 Pt)
4. While paying attention to the nominal variables, propose a linear model of the number of bicycles borrowed per day (variable " cnt ") by keeping only the independent variables significant at 95 %. Comment on the model. (4 pts)
5. By relying on visualization of the residuals, check and comment if the estimation conditions of the model are verified. (1 pts)

Solution

1. point cloud + color for weekdays and weekends. Good correlation with different slope depending on the day
2. sun 4876.786177 + - 1879.483989 cloud 4035.862348 + - 1809.109918

Exercise 2 (statistics and segmentation methods) 40 min (/ 5 pts)
This exercise is based on a set of traffic data collected at a point on a California five-lane freeway, in one direction of traffic, for the day of January 12, 2016 (file `d04_text_404905_raw_2017_02_14.txt`). The data is aggregated at 5 min intervals and the attributes are as follows:

- flow: the number of vehicles
- occupancy: the occupancy rate (proportion of the time that the sensor is occupied by a vehicle)
- speed: the average speed

Please answer the following questions:

1. Describe the distribution of mean speeds by descriptive statistics (1 Pt)
2. Calculate the confidence interval of the mean speeds at 90 and 95 %. (1 Pt)
3. Using an appropriate data mining method, identify groups of traffic conditions from the variables of number of vehicles (variable " flow ") and average speed (variable " speed "). Describe the groups. (3 Pts)

Solution

1. count 288.000000 mean 60.970486 std 18.476275 min 14.500000 first quartile 57.975000 median 71.100000 last quartile 72.300000 max 75.400000
2. 95 (58.82569793207994, 63.115274290142288) 90 (59.174089920552007, 62.766882301670222)

Exercise 3 (spatial analysis, databases and SQL) 45 min (/ 6 pts)

You have the following tables.

- arrondissements table:

Field	Type
id _arround	Integer
name _arround	VARCHAR (255)
Geom	Geometry (MultiPolygon, 32188)

- reseau_routier table:

Field	Type
id_lien_routier	Integer
Geom	Geometry (MultiLinestring, 32188)

- reseau_cyclable table:

Field	Type
id_lien_cyclable	Integer
Geom	Geometry (MultiLinestring, 32188)

Propose a method, for example in the form of an SQL query with spatial functions, in order to determine, by district, the proportion of the road network that contains a cycle lane. The list of spatial functions is presented in the table 1.

Solution
The steps of the method are as follows:

1. Creation of a table of road links by district:
   CREATE TABLE public.reseau_routier_arrond AS SELECT l.*, R.id_arrond, ST_Intersection (l.geom, r.geom) as geom_intersection FROM public.reseau_routier l INNER JOIN public.arrondissements r ON ST_Intersects (l.geom, r.geom);
2. Creation of a table of cycle links by district:
   CREATE TABLE public.reseau_cyclable_arrond AS SELECT l. *, R.id_arrond, ST_Intersection (l.geom, r.geom) as geom_intersection FROM public.reseau_cyclable l INNER JOIN public.arrondissements r ON ST_Intersects (l.geom, r.geom);

Function	Description
ST_Area (g1)	Returns the area of the surface if it is a Polygon or MultiPolygon
ST_Dwithin(g1, g2, distance_of_srid)	Returns true if the geometries are within the specified distance of one another
ST_Intersection (geomA, geomB)	Returns a geometry that represents the shared portion of geomA and geomB
ST_Intersects (geomA, geomB)	Returns TRUE if the Geometries / Geography "spatially intersect in 2D"
ST_Length (g1)	Returns the 2d length of the geometry if it is a linestring or multilinestring
ST_X (g1)	Return the X coordinate of the point
ST_Y (g1)	Return the Y coordinate of the point

Table 1: List of spatial functions

3. Extraction of cycle paths which are at a certain distance from the road network (here 10m):
   CREATE TABLE public.reseau_cyclable_ arrondissement_within10m AS SELECT DISTINCT ON (a.id_link_cyclable) a.* FROM public.reseau_cyclable_rounded a, public.reseau_routier_arrond b WHERE ST_DWithin (a.geom_intersection, b.geom_intersection, 10)

4. Calculation of the lengths of the selected cycle network and of the road network by district:
   CREATE TABLE arrondissements_lengths AS SELECT l.id_round, sum (ST_Length (c.geom_intersection)) / sum (ST_Length (r.geom_intersection)) as percentage_reseau_cyclable FROM arrondissements l LEFT JOIN reseau_cyclable_arrond_within10m c ON l.id_round = c.id_round LEFT JOIN reseau_routier_arrond r ON l.id_round = r. id_round GROUP BY l.name_arron;

Bonus point (/1 pt)
Explain one of the reasons why correlation should not be confused with cause and effect.