Midterm exam

N. Saunier

October 20, 2020

Introduction
In this unusual context of periodic online checks, please:

- note that it has 6 questions counting respectively for 2, 3, 5, 1, 5 and 4 points (the total score is out of 20), with no possible return between the questions, for a duration of two hours;
- note that all documents are allowed;
- pay particular attention to the wording and clarity of the result or answer you give to each question;
- note that some exercises require recently published data on the use of safe active transportation circuits (SATC), put in an SQLite database for the exam exercises available in the exam section;
  - I added in the "pedestrian" and "bikes" tables columns extracted from the counting timestamp column: year ("year" column), month of the year ("month" column), day of week (starting with 0 for Sunday) ("day" column) and time of day ("hour" column);
- note that you can drop files to answer some questions.

Exercise 1 (/ 2 pts)

Propose a method making it possible to collect the counts of pedestrians / cyclists recorded in the "pedestrians" table of the database on SATC. Can this method collect speed data on pedestrians? If not, propose a method to measure the speed of pedestrians.

Solution
Methods for counting pedestrians are: manual counting, the use of an infrared sensor or a video sensor with computer vision algorithms. For cyclists, we can mention manual methods, the use of magnetic loops or video sensors with computer vision algorithms.

Some of these methods do not allow the speed to be measured. The methods operating by detection, if they are doubled, make it possible to determine the speed from the instants of passage. The positions of the users in successive video images make it possible to determine their speed. The only sensor that can directly measure instantaneous speed is radar (and it also allows you to count), although it is more rarely used for these users.

Exercise 2 (/ 3 pts)

The City of Montreal wishes to investigate the motives of pedestrians moving on the SATC. What is the reference population for this survey? What type of survey is it and what survey technique do you recommend?

Solution

- The reference population consists of pedestrians moving on the SATC.
- The type of survey is a section survey, taking place over a period of time, for example a few days or weeks.
- The most natural investigative technique is interception, stopping pedestrians on the SATC to ask them questions.

Exercise 3 (/ 5 pts)

We are interested in creating an information system for collecting counting and speed data on the streets of Montreal. For this, it is necessary to model the different objects and concepts necessary for its operation. These entities are (at a minimum) the following: user, count line, passage, street, sensor, vehicle. A user, possibly in a vehicle, will be counted when he crosses a (virtual) count line by a sensor placed on a certain street.

1. Provide a data model in the form of an Entity / Association diagram involving all the entities listed above. Add attributes (indicating the identifiers) and associations between entities with their cardinalities, minimum and maximum, and functionalities.
2. Translate the Entity / Association schema into a relational schema. Clearly state the primary and outer keys (and what the outer keys refer to), and provide types for the attributes.

The response can be made in the text frame, in a handwritten or handwritten document and photographed.

Solution

1. The entities and their attributes are as follows (the identifier of each entity is in**bold**):

   user id, name, date of birth, gender

   count line id, origin coordinate, end coordinates, installation start time stamp, installation end time stamp

   passage id, timestamp

   street id, name, line (line type vector field), city

   sensor id, type of sensor, positioning coordinate

   vehicle registration number, vehicle type, make, model

   The associations are as follows (it is desirable to name the associations and make a diagram):

- street-count line: a street has had (historically) 0-n count lines, one count line is installed on 1-1 street. The functionality is 1-n.
- count line-sensor: a count line requires 1-1 sensors, a sensor may have been installed on 0-n count lines. The functionality is n-1.
- passage-count line: a passage takes place at 1-1 count line, a count line can count 0-n passages. The functionality is n-1.
- passage-user: a passage implies 1-1 users, a user can make 0-n passages. The functionality is 1-n.
- passage-vehicle: a passage implies 0-1 vehicle, a vehicle can make 0-n passages. The functionality is 1-n.

The count line can be seen as an association between street and sensor (positioning of the sensor to count on a street for a given period). If several users cross the line in the same vehicle, we would make as many passages as users, by linking them to the same vehicle.

2. Each entity becomes a table (Users, CountLines, Passages, Streets, Sensors, Vehicles). It is not necessary to add tables since there are no n-m relationships.The following external keys must be added for the 1-n associations:

   - streetId and sensorId in CountLines to refer respectively to Streets.id and Sensors.id;
   - lineId in Passages to refer to CountLines.id;
   - userId and vehiculeId in Passages to refer to Usagers.id and Vehicules.id respectively.

Exercise 4 (/ 1 pt)

What is the primary key of the pedestrian / bicycle table? Which field (s) should be declared as a foreign key?

Solution

The primary key of the “ pedestrians ” / “ bikes ” table is the composite key (id_station, timestamp_local). The id_station field is a foreign key that refers to the id_station field of the “ stations ” table.

Exercise 5 (/ 5 pts)

Please extract the following information from the SATC database using SQL query. Give the answer (an extract if it is too long) and the SQL query used to get the answer.

1. what is the station with the highest traffic (pedestrians / bicycles) per day in June / July / August / September?
2. what is the SATC (defined by the "name_satc" column of the "stations" table) with the highest traffic (pedestrians / bicycles) per day on weekdays / weekends?
3. for a month of your choice, create a view of the average hourly counts per weekday for each station (with its "name_station" in the result).
4. for a station of your choice (to be specified), calculate the ratio between the two directions of movement according to the days of the week.

5. display the average number of pedestrians / cyclists per weekday and month for stations with an average of pedestrians / cyclists per 15 min period greater than 150/40.

Solution
Here are the queries to get the correct answer:

1. for cyclist ridership in July:
```
SELECT id_station, day, SUM (count_total) FROM bikes
WHERE month = 7 GROUP BY id_station, day
ORDER BY AVG (count_total) DESC
```

2.
```
SELECT stations.name_satc, day, AVG (count_total) FROM bikes,
stations
WHERE (day BETWEEN 1 and 5) AND (bikes.id_station = stations.id_station)
GROUP BY stations.id_station, day ORDER BY AVG (count_total) DESC
```

3. for the month of June:
```
CREATE VIEW profile_time AS
SELECT stations.name_satc, time, AVG (count_total) FROM bikes,
stations
WHERE (month = 6) AND (bikes.id_station = stations.id_station)
GROUP BY stations.id_station, day, time
```

4. for station V07, with the two ways of calculating the average ratio as the average of the ratios or the ratio of the sums:
```
SELECT day, SUM(count_dir_north) / SUM(count_dir_south),
AVG (count_dir_north / count_dir_south)
FROM bikes WHERE id_station = 'V07' GROUP BY day
```

5. (note the subquery)
```
SELECT B.id_station, B.month, B.day, AVG (B.count_total)
FROM BIKES B, (SELECT id_station, AVG (count_total) as n15
FROM bikes GROUP BY id_station) B15
WHERE B15.n15> 40 AND B.id_station = B15.id_station
GROUP BY B.id_station, B.month, B.day
```

Exercise 6 (/ 4 pts)

1. What type of spatial data is used to record the location of stations (raster or vector data)?

2. If you have at your disposal a data layer for the boroughs of Montreal (whose geometry is a polygon) with useful attributes for your analysis, explain the procedure to follow to combine the information of the boroughs with each station.

3. Taking into account the spatial reference system, describe a procedure for calculating the straight line distance between all stations.

4. Discuss the suitability of the geometry chosen for the location of the counting points.

Note that this question does not require any data manipulation.

Solution

1. The location of stations is represented by a point, so it is vector data.
2. A spatial join must be made between the two spatial data layers, which will allow the attributes of the boroughs (zonal data) to be associated with each station (point) according to the borough in which the station is located.
3. The stations must first be converted into an orthonormal coordinate reference system where distances can be calculated, for example the MTM coordinate system. We then follow an algorithm for calculating the Euclidean distance matrix (straight line) between each pair of stations. We initialize a square matrix of size $n \times n$ for the n stations, then with a double loop on the stations, we calculate the distance between each station (see distance calculation algorithm seen in progress).
4. The position of the counting places is represented by a point. Concretely, in the field, it is unusual and impractical to count road users passing over a point. We generally consider a (virtual) row, which could be represented as such in the database.