Contrôle périodique

N. Saunier

19 octobre 2021

Veuillez

- noter le barème (la note totale est sur 20) et le temps indicatif à consacrer à chaque exercice;
- indiquer clairement les numéros des questions que vous traitez et vos réponses correspondantes (et souligner ou encadrer les résultats numériques);
- apporter une attention particulière à la rédaction et à la définition des notations que vous employez;
- noter que certains exercices nécessitent des fichiers disponibles sur Moodle (Section "Contrôle périodique").

Exercice 1 (collecte de données) 45 min (/7.5 pts)

1. Nommer et décrire deux contraintes pouvant limiter le processus d'acquisition ou de collecte de données. (1 pt)
2. Deux situations hypothétiques sont présentées ci-dessous. Pour chaque situation, veuillez (2 pts)
   - déterminer la caractéristique de la méthode de collecte de données qui sera problématique au moment du traitement et de l'analyse des données;
   - déterminer si la problématique est liée à la variabilité des observations, à un biais dans les données ou à la facilité de traitement des données.

   (a) Une enquête à bord est réalisée afin de dresser le portrait de la mobilité de l'ensemble des utilisateurs des trains de banlieue du Réseau de transport métropolitain (RTM, exo). À cet effet, un questionnaire a été construit de manière rigoureuse, et les enquêteurs ont reçu une formation pour éviter d'influencer les réponses des personnes enquêtées. Pour les besoins de l'enquête, un échantillon important est souhaité afin d'augmenter la précision. Exo décide donc d'enquêter l'ensemble des utilisateurs de la ligne de train vers Mont-Saint-Hilaire, une des lignes les plus achalandées du réseau, pendant un jour de semaine type de novembre. Les questionnaires sont remplis par les enquêteurs, directement sur une tablette à partir des réponses des usagers.

   (b) Un boulevard majeur est congestionné pendant les heures de pointe. Pendant cette période, un nombre important de déplacements semblent être réaffectés sur une rue locale et résidentielle parallèle au boulevard.

Après plusieurs plaintes de résidents, vous êtes mandaté pour caractériser le transit dans le secteur et pour évaluer le taux de transit aux périodes de pointe. Vous faites donc appel à un fournisseur de données provenant de cellulaires (données des déplacements géolocalisés) pour évaluer la provenance et la destination de l'ensemble des déplacements réalisés sur la rue locale.
L'échantillon est évalué à 10 % de l'ensemble des véhicules étant passés par la rue locale et les données sont fournies pour chaque jour de la semaine sous forme de moyenne de l'ensemble de l'année 2020. Les données du fournisseur ont le format suivant:

Id	Origine	Destination	Jour	Nombre d'observations
1	Rue A	Rue B	lundi	54
2	Rue A	Rue B	mardi	48
...				
326	Rue C	Rue D	lundi	60
...				

3. Vous effectuez un relevé de vitesses par radar pour déterminer la vitesse moyenne sur une route à l'heure de pointe. Quelle est la taille minimale de l'échantillon que vous devez recueillir pour déterminer la vitesse moyenne sur l'axe si vous désirez un niveau de confiance de 95 % et que vous souhaitez absolument respecter une marge d'erreur (tolérance) de 3 km/h au maximum? Vous savez que le débit de la route étudiée à l'heure de pointe est de 1300 véh/h. Plusieurs études similaires réalisées précédemment sur des axes semblables montrent que l'écart-type des vitesses varie d'une route à l'autre entre 5 et 14 km/h. (1.5 pts)

4. Pour la même route que la question précédente, on désire aussi connaître la proportion de camions: quelle est la taille de l'échantillon nécessaire pour la déterminer avec une marge d'erreur (tolérance de 4 %) avec un niveau de confiance de 95 %? Quelle est la taille de l'échantillon nécessaire pour relever les vitesses et le type de véhicule en une seule collecte? (1.5 pts)

5. Donner la définition de la population de référence et de la base de sondage? Donner un exemple le même ensemble peut jouer les deux rôles. (1.5 pts)

Solution

1. Le cours mentionnait cinq contraintes pour la collecte de données:
   - coût
   - respect de la vie privée
   - disponibilité d'une méthode de collecte
   - propriété des données
   - stockage

2. Voici des éléments de réponse aux questions:
   - La couverture spatiale est déficiente (enquête sur une seule ligne du RTM, alors que l'enquête vise l'ensemble des utilisateurs) et entraînera un biais dans les résultats.

- La résolution temporelle des données utilisées provenant de cellulaires est déficiente (l'étude vise les heures de pointe, mais les données fournissent des données pour une année entière sans spécifier les heures des déplacements) et entraînera un biais dans les analyses qui ne peut être corrigé par un traitement des données.

3. La taille minimale de l'échantillon est de $n = \frac{k^2_{\alpha/2}\sigma^2}{e^2} = \frac{1.96^2 14^2}{3^2} = 83.7 \approx 84$ (en prenant le pire des cas pour l'écart-type des vitesses).

4. La taille minimale de l'échantillon est de $n = \frac{k^2_{\alpha/2}p(1-p)}{e^2} = \frac{1.96^2 0.5(1-0.5)}{0.04^2} = 600.2 \approx 601$ (en prenant $p = 0.5$ pour le pire des cas). Il suffit donc de collecter les informations de vitesse et de type (camion ou pas) pour 601 véhicules pour avoir la précision voulue sur les moyennes des deux attributs (vitesse et type de véhicule).

5. La population de référence est l'ensemble pour lequel nous cherchons à obtenir des informations. Une base de sondage est la population utilisée pour tirer un échantillon (sous-ensemble) qui sera enquêté, et qui peut être différente de la population cible selon les sources de données disponibles. La base de sondage est la population de référence par exemple lorsque les abonnés d'un service sont enquêtés et qu'on utilise la base de données des abonnés comme base de sondage.

Exercice 2 (base de données et SQL) 30 min (/4 pts)

Télécharger la base de données `ReseauRoutier.db` de la section du contrôle périodique sur Moodle qui décrit un réseau routier, contenant trois tables pour des carrefours (table "nodes"), des segments routiers (table "sections") et les virages permis aux carrefours (table "turnings"). Indiquer les requêtes SQL qui permettent de répondre aux questions suivantes (il n'est pas demandé de donner le résultat de la requête). Le logiciel "DB Browser for SQLite" est disponible sur les ordinateurs si vous désirez tester vos requêtes.

1. Calculer la limite de vitesse (champ "speed") moyenne sur les sections de l'ensemble du réseau, sans pondération, puis pondérée par la longueur (champ "length") des sections. (1 pt)

2. Calculer et afficher le nombre de sections pour chaque classe de limite de vitesse. (1 pt)

3. Afficher l'ensemble de la table des virages permis ("turnings") et ajouter l'attribut qui définit le type de carrefour (champ "nodetype" de la table "nodes") afin de préciser le type d'intersection dans lequel le mouvent tournant se situe. (1 pt)

4. Afficher de manière décroissante les attributs des carrefours en fonction du nombre de mouvements tournants soumis à un panneau d'arrêt (sachant qu'un mouvement géré par un arrêt est représenté par la valeur 2 du champ "sign"). La liste doit comprendre l'identifiant unique du carrefour, le nom du carrefour (champ "name") et le nombre de mouvements gérés par un panneau d'arrêt. (1 pt)

Solution

1. `SELECT AVG(speed), SUM(speed*length)/SUM(length) FROM sections`

2. `SELECT speed, count(*) FROM sections GROUP BY speed`

3. ```
   SELECT turnings.*, nodes.nodetype FROM turnings, nodes
   WHERE turnings.id_node = nodes.id
   ```

4. ```
   SELECT id_node, name, COUNT(turnings.id) as nturnings
   FROM turnings, nodes
   WHERE turnings.id_node = nodes.id AND turnings.sign=2
   GROUP BY id_node
   ORDER BY nturnings DESC
   ```

Exercice 3 (modèle de données et données spatiales) (/6.5 pts)

Nous voulons concevoir un modèle de données pour l'enquête sur les déplacements effectuée par la Ville de Montréal à l'aide de l'application mobile MTL trajet. Après avoir installé l'application sur leur téléphone, les participants répondent à un premier questionnaire sur leurs caractéristiques socio-démographiques et leurs habitudes de transport. L'application enregistre ensuite leurs déplacements (capteur GNSS) pendant 30 jours. Pour chaque déplacement, lorsqu'elle détecte la fin du déplacement, l'application demande au répondant des informations complémentaires comme le mode et le motif du déplacement.

1. Proposer un modèle pour les données collectées avec l'application mobile MTL trajet sous forme d'un diagramme Entité/Association impliquant au minimum les entités suivantes: participant, déplacement, point GNSS. Ajouter des attributs (incluant l'identifiant) et les associations entre entités, avec leurs cardinalités minimale et maximale, et les fonctionnalités. (1.5 pts)

2. Traduire le schéma Entité/Association en schéma relationnel. Indiquer clairement les clefs primaires et externes, et proposer des types pour les attributs. (1 pt)

3. Discuter à quel type de données spatiales (matricielles ou vectorielles) les données GNSS correspondent, avec deux avantages et deux inconvénients de ce type de données. Donner un exemple de données de l'autre type pertinent pour le transport. (2 pts)

4. Proposer un système de référence des coordonnées adapté aux données collectées par l'application MTL Trajet et justifier le choix. (1 pt)

5. Pour tirer partie des fonctionnalités d'un système d'information géographique (SIG), on désire utiliser une base de données spatiales pour enregistrer les déplacements. Discuter comment modifier le modèle de données en utilisant des types de données spatiales pour permettre d'afficher directement les déplacements des usagers dans un SIG. (1 pt)

Solution

1. Les entités et leurs attributs sont les suivants (l'identifiant de chaque entité est en **gras**):

   participant id, nom, date de naissance, occupation, code postal de résidence

   déplacement id, mode, motif

point GNSS id, date, heure, coordonnée x, coordonnée y

Les associations sont les suivantes (il est souhaitable de nommer les associations et de faire un schéma):

- participant-déplacement: un participant fait 0-n déplacements, un déplacement est fait par 1-1 participant. La fonctionnalité est 1-n.
- déplacement-point GNSS: un déplacement est constitué de 1-n points GNSS (on pourrait probablement dire 2 à n), un point GNSS fait partie de 1-1 déplacement. La fonctionnalité est 1-n.

2. Chaque entité devient une table (Participants, Déplacements, Points GNSS). Il n'est pas nécessaire d'ajouter de tables pour représenter les associations n-m. Il faut ajouter des clefs externes suivantes pour les associations 1-n:

   - participantId dans Déplacement faisant référence à Participants.id;
   - déplacementId dans Points GNSS faisant référence à Déplacements.id.

   Voici quelques types de données pour les attributs: le nom, occupation, code postal, mode et motif sont des attributs catégoriels, représentés par du texte. La date de naissance est de type date (comme date de la table Points GNSS). Les coordonnées x et y sont des nombres décimaux, ou du texte selon le système de référence spatial utilisé.

3. Les capteurs GNSS mesurent les positions successives des usagers sous forme de série de points. Les points sont des données vectorielles. Les avantages et inconvénients sont discutés dans les notes de cours. Un exemple de données matricielles pertinent pour le transport est les données de modèle numérique de surface, utilées pour la conception des routes et les pentes dans les calculs de chemins par exemple pour le vélo.

4. On pourrait utiliser les sytèmes de coordonnées MTM ou UTM (plus précis en longitude), où la zone de Montréal est respectivement 18 et 8.

5. Le plus simple consisterait à remplacer les champs des coordonnées x et y dans la table des points par un champ point. Une alternative consisterait à supprimer la table des points GNSS et à simplement avoir un champ de type ligne dans la table Déplacement. Un enjeu serait de représenter le temps, sans bonne solution si les points ne sont pas enregistrés à intervalle de temps régulier.

Exercice 4 (traitement de données) 30 min (/2 pts)

1. On dispose de données de véhicules flottants se déplaçant sur le réseau pour la collecte de temps de parcours sur des trajets et on enregistre leurs positions longitudinales successives (distance parcourue en fonction du temps) à intervalle de temps régulier. Indiquer la sortie de l'algorithme suivant (ce qui est calculé): (1 pt)

   entrée: n positions longitudinales $d_1, ..., d_n$, durée Δt entre les enregistrements de position
   sortie: ?

début
$x = 0$
pour $i = 2...n$
$x = x + \frac{d_i - d_{i-1}}{\Delta t}$
renvoyer $\frac{x}{n-1}$
fin

2. Modifier l'algorithme précédent de sorte qu'il mesure la durée de temps passé par le véhicule à l'arrêt. (1 pt)

Solution

1. Voici une solution possible:

   - Entrées: une série (liste) de n positions (une position est un point, qui est une liste de dimension 2), des réseaux de transport (un graphe par réseau de transport), une carte de points d'intérêt (lieux de travail, magasins, restaurants, café, etc.)

   - Sortie: une liste de taille $n - 1$ du mode utilisé pour se déplacer entre chaque position

2. L'algorithme présenté calcule la vitesse moyenne du déplacement.

3. Voici une solution (noter qu'on pourrait ajouter une petite tolérance pour le test de non déplacement (d_i égal à d_{i-1}), comme un seuil de distance maximale lié au bruit des positions GNSS en l'absence de mouvement):

   entrée: n positions longitudinales $d_1, ..., d_n$, pas de temps Δt
   sortie: durée de temps à l'arrêt

   début
   $duree_{arret} = 0$
   pour $i = 2...n$
   si d_i égal à d_{i-1}
   $duree_{arret} = duree_{arret} + \Delta t$
   renvoyer $duree_{arret}$
   fin