Description des problèmes soumis au 12e ARPI de Montréal

Air Canada
Construction d'un programme de maintenance

Air Canada doit assurer la maintenance de ses aéronefs grâce à un programme précis de maintenance en considérant l'accumulation des heures, cycles ou jours sur ses actifs depuis leur dernière remise à neuf. Accessoirement des défauts se présentent aussi au fil du temps et doivent être corrigés dans le cadre de contraintes semblables (basées sur les heures, cycles ou jours). L'équipe de maintenance doit construire un plan de maintenance incluant les détails de chaque journée sans que le rendement en soit affecté, en plus de corriger les défauts aussitôt que possible sans dépasser les échéances liées aux limites sur les nombres d'heures, de cycles ou de jours ; ce plan doit tenir compte des ressources limitées mises à sa disposition pour effectuer ces tâches à de nombreux sites.

L'objectif de ce problème est de construire un programme de maintenance qui tienne compte de plusieurs critères et contraintes : minimiser le rendement qui reste à une date maximale d'utilisation pour un actif donné, maximiser le rendement qui reste pour un défaut survenu entretemps et prendre en compte des bornes supérieures pour les ressources disponibles à de nombreux sites.

Banque Nationale du Canada
Métriques pour la performance, la sécurité et la confidentialité d'une solution d'anonymisation

La Banque Nationale du Canada (BNC) souhaite créer toujours plus de valeur pour ses clients grâce aux données, tout en protégeant ces données et empêchant toute utilisation illicite qui amènerait un bris de confiance. Les chercheurs en protection de vie privée parlent de compromis entre vie privée et utilité. En particulier la sécurité et la confidentialité des données des clients sont des priorités pour la BNC. Ces dernières années, afin d'atteindre ses objectifs en termes de vie privée et d'utilité et de renforcer sa position en matière de sécurité et de confidentialité des données, la BNC a consacré des ressources au développement de ses capacités d'anonymisation des données. Toutefois évaluer avec précision l'impact de chaque nouvelle capacité d'anonymisation est encore un défi, tant pour la BNC que pour les scientifiques travaillant dans ce domaine.

La BNC souhaiterait disposer d'une méthodologie robuste afin de mesurer les risques présentés par l'application d'une méthode d'anonymisation, quels que soient le modèle sous-jacent à cette méthode et l'algorithme qu'elle emploie.

Voici trois risques importants encourus lorsqu'on souhaite publier des données dites désensibilisées : la réidentification, l'appartenance et la divulgation d'attributs. Dans la **réidentification**, une personne malveillante arrive à identifier un certain nombre d'individus en utilisant les données publiées et possiblement des données auxiliaires. Dans l'**appartenance**, une personne malveillante peut réussir à prouver qu'un individu a fait partie de la collecte des données publiées. Dans la **divulgation d'attributs**, une personne malveillante peut reconstruire des informations personnelles à partir des données publiées.

Étant donné ces risques, quelles sont les meilleures métriques pour mesurer la performance, la sécurité et la protection de la vie privée d'une méthode d'anonymisation déployée dans le contexte d'un processus d'affaires d'exploitation des données ?

La BNC s'attend à ce que les membres de l'équipe proposent une liste de métriques permettant de mesurer les risques associés à la publication de données anonymisées. Ces métriques doivent être indépendantes de la méthode d'anonymisation choisie. Une meilleure compréhension des compromis entre les propriétés mesurées par ces métriques et l'utilité des données anonymisées serait la bienvenue. Les membres de l'équipe auront à analyser un algorithme d'anonymisation afin de mesurer sa performance.

Une paire de jeux de données de type transactionnel sera fournie au début de l'atelier. Un des jeux sera de source publique et contiendra des données identifiantes. L'autre jeu de données proviendra de la même source que le premier, mais ses données auront été anonymisées. Vingt-quatre heures avant la fin de l'atelier, une nouvelle paire de jeux de données, provenant d'une autre source, sera fournie aux participants pour qu'ils valident leurs analyses. Les analyses devront donc être assez générales pour fonctionner sur d'autres paires de jeux de données.

Beneva

Modèles de survie et données incomplètes

Les modèles de survie sont souvent utilisés pour évaluer le temps pendant lequel les clients demeureront actifs, sans interruption, dans leurs relations avec une entreprise. Dans ce contexte l'utilisation des modèles de survie présente quelques difficultés. Des articles scientifiques ont proposé des solutions à plusieurs de ces difficultés, comme la censure à droite ou à gauche et les covariables qui varient dans le temps. Dans la pratique, on rencontre quelques problèmes supplémentaires. En voici un. Dans certains cas, nous connaissons la date de commencement d'un client dans l'entreprise, mais les informations sur les covariables ne sont pas disponibles pour toute la période où le client a eu des relations avec l'entreprise. Par exemple, un client peut être en relation avec une entreprise depuis 2002 (cette date est connue), mais les valeurs de ses covariables (par exemple les produits détenus) ne sont connues que depuis 2010. L'accès à l'historique de ce client est donc limité. Les covariables peuvent être modifiées au fil du temps mais seules les modifications ayant eu lieu depuis 2010 sont connues. On ne connaît pas leurs valeurs avant 2010.

Dans ce cadre plusieurs questions se posent. Quelle est la meilleure façon d'exploiter les informations sur ce client dans un modèle de survie (forêt de survie, modèles de Cox, ou autre) afin d'inclure toute la durée de survie du client bien que le suivi des covariables soit limité? Quels sont les inconvénients à retirer de l'échantillon ce type de clients ? Bien que le but de ce problème soit de nature méthodologique ou théorique, Beneva fournira un jeu de données synthétique représentant les véritables données. Ceci permettra aux membres de l'équipe de tester leurs propositions de solutions.

Environnement et Changement Climatique Canada (ECCC)
Prototype d'un modèle hybride statistique / dynamique pour des simulations hydrodynamiques à haute résolution

ECCC met au point des outils opérationnels de prévision environnementale permettant de caractériser et de suivre l'évolution de différentes composantes du Système Terre, et ce, 24h sur 24 et 7 jours sur 7. Ces systèmes de prévision comprennent typiquement une cascade de modèles couplés représentant les conditions atmosphériques, océaniques, de glaces, hydrologiques, hydrodynamiques et écosystémiques, de l'échelle globale jusqu'à l'échelle régionale, voire locale. Dans les lacs et rivières, ces systèmes sont utilisés comme outils intégrés de support à la prise de décision et à la gestion des eaux, comme aide à la navigation, en support aux opérations de recherche et de sauvetage et en réponse aux urgences environnementales.

Plusieurs applications requièrent la modélisation à haute résolution (spatiale et temporelle) de variables critiques comme la profondeur de l'eau ou la vitesse des courants, notamment en régions côtières ou à proximité des infrastructures. Malgré la disponibilité de certaines données de base à haute résolution et la puissance de calcul grandissante des superordinateurs, les simulations à haute résolution demeurent très coûteuses et les temps de calcul requis sont parfois inadéquats, à cause de la rapidité de réponse requise dans certaines situations urgentes. Par conséquent un compromis est généralement adopté, tant au niveau de la résolution spatiale que de l'échéance des prévisions (habituellement de l'ordre de quelques jours).

Une solution hybride (statistique / dynamique) est recherchée afin d'élargir le champ d'applicabilité des modèles en augmentant, à coût réduit, la rapidité de calcul et la résolution spatiale des variables hydrodynamiques simulées. Une telle solution rendrait également possible des prévisions (déterministes et ensemblistes) à plus longue échéance ou des projections climatiques.

Pour accomplir cette tâche, plusieurs émulateurs, ou modèles de remplacement, peuvent être mis en place, utilisant des approches empruntées à l'intelligence artificielle et des méthodes statistiques ou numériques visant à réduire la dimensionnalité du problème simulé. Ces techniques permettraient, par exemple, de générer des champs à haute résolution à partir de simulations à basse résolution. Dans le cadre du présent atelier, les objectifs sont (1) de recenser les approches existantes et d'élaborer une stratégie pour l'application d'émulateurs dans un contexte de prévision hydrodynamique pour lacs et rivières, et (2) de mettre en place les bases d'un prototype simple d'émulateur permettant d'établir une relation statistique / dynamique entre deux modèles existants.

ECCC dispose de champs géophysiques et de simulations hydrodynamiques à basse et haute résolutions sur différents domaines, notamment sur le Lac Érié, le fleuve Saint-Laurent, le Lac Champlain et le port de Saint-John (au Nouveau-Brunswick). Des résultats de calcul basés sur les modèles seront mis à la disposition des membres de l'équipe pour l'un de ces domaines, afin d'appuyer leur travail.

Hydro-Québec (TransÉnergie et Équipement)
Prévision de la demande par poste

Contexte

Hydro-Québec TransÉnergie et Équipement (HQTÉ) cherche à prévoir comment un groupe d'équipements appelé poste satellite de moyenne tension va alimenter en électricité son secteur de distribution pendant les prochaines heures et les prochains jours, étant donné les conditions météorologiques et d'autres paramètres prévisionnels. Le fonctionnement d'un poste est caractérisé par plusieurs variables dépendantes, mesurées en différentes unités (puissance, intensité, etc.). Leur dépendance provient du fait qu'elles sont associées aux sorties d'un ou plusieurs équipements reliés entre eux (transformateur, disjoncteur, etc.).

Problème

HQ veut construire un modèle unique qui, par apprentissage sur des mesures observées, va simuler dans un premier temps l'ensemble des variables dépendantes, expliquées à partir des variables indépendantes ou intrants (température, vent, enneigement, précipitations, jour de la semaine, date, heure, profils correcteurs par type de client, niveau de charge macroscopique). Les variables dépendantes ou extrants sont la puissance active, la puissance réactive et l'intensité aux points de sortie.

Solution recherchée

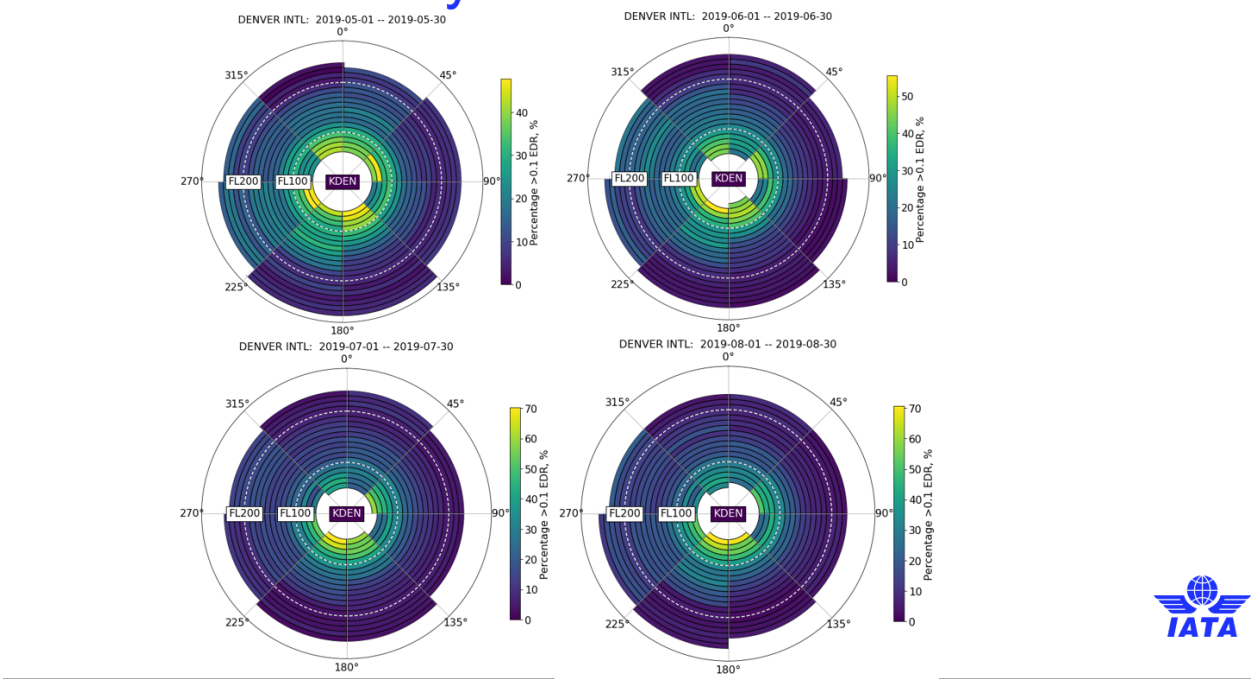
Ce modèle pourra être confronté à un ensemble de modèles existants (chacun dédié à une variable dépendante), afin d'identifier la pertinence des gains provenant de l'exploitation de leur dépendance par ce modèle unique. Des mesures statistiques telles que la moyenne absolue des écarts sur les résidus pourront être utilisées à cette fin. Selon les résultats, il sera envisageable de partitionner ou séquencer le calcul de cet ensemble de variables. Ensuite le même exercice pourra être réalisé avec des données prévisionnelles pour les intrants afin d'en mesurer la fiabilité dans un contexte opérationnel. HQTÉ fournira des données horaires pour un horizon de deux ans.

International Air Transport Association (IATA)
Turbulence atmosphérique : carte de chaleur et diagramme saisonnier

Dans ce projet nous considérons deux questions : celle de construire une carte de chaleur pour la turbulence atmosphérique (abrégée turbulence ci-après) et celle de construire un diagramme saisonnier. Le première question consiste à construire une carte de chaleur pour un niveau de vol donné, plus ou moins deux mille pieds. Par exemple on aimerait voir la carte de chaleur de la région concernée pour le niveau de vol 24, ce qui veut dire que toutes les données entre 22 mille et 26 mille pieds seraient prises en compte. IATA peut fournir deux types de jeu de données : un jeu pour une période de quatre (4) heures et un jeu de données historiques, correspondant à plusieurs mois de données. L'IATA aimerait que les membres de l'équipe lui donnent des conseils

sur l'observation de tendances, en utilisant des outils tels que la statistique et l'apprentissage machine.

La deuxième question consiste à construire un diagramme saisonnier pour un aéroport donné, mettant en évidence la direction, l'intensité et le niveau de la turbulence (voir la figure ci-dessous). Étant donné un jeu de données de 12 mois pour la région incluant l'aéroport, les membres de l'équipe devront construire un diagramme (ou des diagrammes) illustrant les caractéristiques de la turbulence dans cette région. L'intensité de la turbulence par élévation pour chaque secteur (pour des intervalles de deux mille



pieds) doit apparaître sur le diagramme. Dans la figure ci-dessus, chaque segment représente un secteur et deux mille pieds : sa couleur représente l'intensité, égale au taux moyen de dissipation des tourbillons (EDR, en anglais), au taux médian ou au pourcentage de ce taux au-dessus d'un certain seuil. Le diagramme doit aussi indiquer la direction de la turbulence autour de l'aéroport, par secteurs de 45 degrés. Il faut construire un diagramme pour chaque saison et intervalle de temps.

De tels diagrammes fournissent au pilote, lors du décollage ou de l'atterrissage, l'intensité attendue de la turbulence par niveau de vol et direction. En outre les membres de l'équipe essaieront de donner des informations sur l'aspect saisonnier de la turbulence pour un aéroport donné. Par exemple, pour l'aéroport de Denver, la probabilité de turbulence lorsque le pilote se prépare à l'atterrissage est de x% à partir du nord pendant l'été. L'IATA veut s'appuyer sur l'expertise des membres de l'équipe pour construire des diagrammes aussi clairs que possible permettant aux pilotes de prendre des décisions sans une avalanche de données.

Radio-Canada
Simplification textuelle automatique pour un diffuseur public

CBC / Radio-Canada a pour mandat de renseigner, d'éclairer et de divertir tous les Canadiens / Canadiennes. De plus, lors de la rédaction de notre plan sur l'équité, la diversité et l'inclusion pour 2022-2025, CBC / Radio-Canada s'est engagé à ce que toutes les personnes vivant au Canada se sentent valorisées, reconnues et entendues par leur diffuseur public d'un océan à l'autre. Radio-Canada publie sur son site internet (radio-canada.ca) entre 450 et 600 articles en moyenne par jour. Ces publications, majoritairement composées de textes, traitent pour la plupart de sujets d'actualité particulièrement complexes (crise sanitaire, climatique, économique, polarisation de la société, conflits internationaux, etc.). Comme une bonne compréhension des enjeux actuels est non seulement nécessaire mais capitale pour participer au débat démocratique, Radio-Canada pense que ses contenus seraient mieux compris par un plus grand nombre de citoyens si des techniques telles que l'*Automatic Text Simplification* (ATS) étaient utilisées. Simplifier et/ou résumer automatiquement certains de ses contenus écrits permettrait de favoriser leur compréhension et de les rendre plus attrayants pour (par exemple) des personnes ayant un niveau de littératie insuffisant, des personnes en situation de neurodiversité, des nouveaux arrivants, et ainsi de suite.

En avril 2021, CBC / Radio-Canada a créé Mauril2, une plateforme d'apprentissage numérique du français et de l'anglais à partir de contenus vidéos et audios du diffuseur public. L'équipe de développement réfléchit actuellement à élargir l'offre d'apprentissage à nos contenus écrits par l'entremise d'une nouvelle activité de compréhension écrite. Pour mettre en place cette nouvelle activité, Radio-Canada aura besoin de la simplification textuelle automatique afin de produire des textes plus simples, adaptés au niveau de compétence des apprenants débutants, à partir de nos contenus originaux. Nous viserons à la fois les apprenants du français et de l'anglais. La simplification textuelle automatique consiste à réduire la complexité d'un texte (en termes de lexique et de syntaxe) tout en conservant son sens original, afin d'améliorer sa lisibilité et sa compréhension. Les méthodes de simplification de phrases se divisent en deux grandes familles, les systèmes modulaires (qui effectuent des opérations de simplification lexicales et/ou syntaxiques de façon itérative et/ou récursive) et les systèmes de bout en bout, inspirés des systèmes de traduction automatique (neuronale), qui apprennent à effectuer plusieurs modifications à la fois à partir de données étiquetées. Dans les systèmes modulaires, la simplification textuelle est le plus souvent réalisée par des transformations effectuées au niveau de la phrase, telles que le remplacement (simplification lexicale), la réorganisation (simplification syntaxique) et le fractionnement.

En participant à l'atelier, l'équipe de Radio-Canada désire explorer les avenues les plus prometteuses dans le domaine de la simplification de textes, particulièrement dans le contexte qui est le sien, celui d'un radiodiffuseur public désireux de rejoindre le plus de citoyens possible pour mettre à leur disposition une information de qualité.

Revenu Québec
Détection d'entreprises à risque

Revenu Québec désire se donner un outil de détection d'entreprises à risque (quelquefois appelées entreprises frauduleuses) basé sur certains travaux déjà publiés. L'article *Guilt-by-Constellation : Fraud Detection by Suspicious Clique Memberships* ([GbC], voir citation ci-dessous) propose une méthode de détection d'entreprises à risque utilisant les particularités d'un graphe biparti où un ensemble d'entreprises est lié à un ensemble de ressources. Voici les étapes de cette méthode : création d'un graphe avec des poids sur les sommets et les arêtes, détection des cliques du graphe, évaluation du score de chaque clique, calcul d'attributs se basant sur les cliques pour chaque entreprise et utilisation des attributs calculés dans un modèle de prédiction.

Cet article fait suite à un autre article, intitulé *GOTCHA! Network-based Fraud Detection for Social Security Fraud* ([GOTCHA], voir citation ci-dessous), proposant également une méthode de détection d'entreprises à risque mais ne faisant pas appel à la notion de clique. Des travaux relatifs à cet article (et presque terminés) ont été réalisés à Revenu Québec et ont consisté à créer un graphe pondéré, à calculer la majorité des attributs et à formuler un modèle de prédiction des entreprises à risque.

Dans ce problème on se préoccupe en particulier du calcul des attributs basés sur les cliques (voir la Section 4.4 de [GbC]). Naturellement il faut détecter les cliques avant de calculer ces attributs. Dans la Section 4.3 du même article, les auteurs utilisent une approche ascendante (« bottom-up ») pour trouver toutes les cliques dans le graphe biparti. Cette approche n'est pas performante et ne peut être utilisée sur des graphes bipartis contenant des dizaines de milliers de sommets entreprises et des millions de sommets ressources. Le but de l'équipe étudiant ce problème est donc de proposer une approche performante pour la détection de cliques. Toutefois, en fin de compte, c'est la liste des attributs basés sur les cliques qui doit être calculée. Revenu Québec aimerait disposer d'une solution globale prenant un graphe pondéré en entrée et produisant une telle liste pour chaque entreprise.

Afin de permettre à l'équipe de tester les algorithmes proposés, Revenu Québec lui fournira des graphes avec des poids sur les sommets et les arêtes ayant les caractéristiques des graphes traités au ministère.

[GbC] Van Vlasselaer, Véronique & Akoglu, Leman & Eliassi-Rad, Tina & Snoeck, Monique & Baesens, Bart. (2015). Guilt-by-Constellation: Fraud Detection by Suspicious Clique Memberships. 2015. 918-927. 10.1109/HICSS.2015.114.

[GOTCHA] Van Vlasselaer, Véronique & Eliassi-Rad, Tina & Akoglu, Leman & Snoeck, Monique & Baesens, Bart. (2016). GOTCHA! Network-based fraud detection for social security fraud. Management Science. 63. 10.1287/mnsc.2016.2489.