Description of the problems submitted to the 12th Montreal IPSW

Air Canada
Building a maintenance plan

At Air Canada Maintenance we must maintain our aircraft through a defined maintenance program using accumulated hours/cycles and/or calendar days on our assets since their last refurbishment. We incidentally also have defects that occur along the way and naturally we must fix those within similar constrained thresholds of hours/cycles or calendar days. We must plan the maintenance program items down to the day of execution without losing yield, whereas we have to plan to rectify the defects as soon as possible without overrunning the due dates derived from the limits on the hours/cycles or calendar days, while accounting for a limited number of resources available to perform the work required at multiple locations.

The objective of this problem is to build a forecasted maintenance plan that combines competing goals: minimize the yield remaining at a maximum usability date for a given asset and maximize the yield remaining for a defect that has been generated along the way. Upper bounds on available resources at multiple stations must also be taken into account.

Banque Nationale du Canada / National Bank of Canada
Measuring the performance, security, and confidentiality of an anonymization approach

The National Bank of Canada wishes to create more value for its clients through data mining, while protecting the clients' data and preventing the illicit use of these data (leading to a breakdown in trust). Researchers investigating the protection of private data have to consider trade-offs between privacy and utility. In particular the security and confidentiality of its clients' data have a high priority for the bank. In recent years, in order to achieve its privacy and utility objectives and strengthen its position in matters of data security and confidentiality, the bank has devoted resources to the development of its data anonymization capacity. The evaluation of a new capacity of this sort, however, is still a challenge for the bank and the researchers working in this field.

The bank would like to develop a robust methodology for measuring the risks involved in the application of an anonymization method. This methodology should be independent from the model underlying the method and the algorithm used.

Here are three important risks that the publication of anonymized data entails: reidentification, inclusion detection, and attribute divulgation. In **reidentification** an attacker succeeds in identifying certain individuals by using published data and possibly other data. In **inclusion detection** an attacker can succeed in proving that an individual is included in the published data. In **attribute divulgation** an attacker is able to recover personal information by using published data.

Given these risks, what are the best metrics for measuring the performance, security, and privacy protection of an anonymization method deployed within a business process involving data mining?

The bank expects the team members to propose a list of metrics for measuring the risks involved in the publication of anonymized data. These metrics should not depend upon the anonymizing method. A better understanding of the trade-offs between the measured risks and the utility of anonymized data would be much appreciated. The team members will have to analyze an anonymization algorithm in order to assess its performance.

A pair of transactional data sets will be provided at the beginning of the workshop. The source of one of the data sets will be a public one and it will contain identifying data. The other data set will have an identical source but its data will be anonymized. Twenty-four hours before the end of the workshop, a new pair of data sets, coming from a different source, will be made available to team members for validation purposes.

Beneva
Survival models and incomplete data

Survival models are often used to evaluate the time clients remain in a relationship with a company, without interruption. Within this context the use of survival models entails some difficulties, such as right-censored or left-censored data or time-varying covariables. In practice still more difficulties are encountered. Here is one of them. In some cases we know the "start date" for a client's involvement but the information on covariables is not available for the entire involvement period. For instance the start date may be 2002 (a known information) but the values of the covariables (e.g., insurance products) are available only for the period beginning in 2010. The client's history is therefore only partially known. Covariables may change over time but only changes occurring in 2010 or later are recorded. Their values before 2010 are not known.

Several questions arise within this context. What is the best way to "mine" the client's informations within a survival model (survival forest, Cox models, etc.) in order to include the entire survival period, in spite of the fact that the covariables information is not complete? What are the drawbacks if this type of client is removed from the sample? Although the proposed problem is theoretical in nature, Beneva will provide the team with a synthetic data set similar to the real data set. This will enable the team members to test the solutions they propose.

Environment and Climate Change Canada
Building a prototype for a statistical / dynamical high-resolution model to be used in hydrodynamical simulations

Environment and Climate Change Canada (ECCC) is developing environmental forecasting tools for characterizing and following the evolution of different components of System Earth, 24 hours a day and 7 days a week. These forecasting tools typically include a cascade of coupled models representing the atmospheric, oceanic, ice-related, hydrologic, hydrodynamic, and ecosystemic conditions, from the global scale to the regional (even local) scale. In lakes and rivers, these systems are used as tools for decision-making and water management, for help in navigation, for supporting search and rescue operations, and for responding to environmental emergencies.

Several applications require the modelling, with a high (spatial and temporal) resolution, of critical variables such as water depth or current speed, especially in coastal regions or in the vicinity of infrastructures. Although some data with varied resolutions are available and the computing power of supercomputers increases all the time, high-resolution simulations are still very expensive and computing times are sometimes inadequate, given that some urgent situations require a quick response. Hence one has to compromise when choosing a spatial resolution and/or a forecasting horizon (usually a few days).

One is looking for a hybrid (statistical / dynamical) solution, in order to enhance the applicability of models by increasing the computing speed and the spatial resolution of simulated hydrodynamic variables. Such a solution would also allow longer-term deterministic or ensemble forecasting or climate projections.

To achieve this several emulators, or replacement models, can be built, using artificial intelligence and statistical or numerical methods to reduce the simulated problem dimensionality. These techniques could allow, for instance, the generation of high-resolution fields from low-resolution simulations. Within the workshop the goals are (1) to survey the existing approaches and devise a strategy for the application of emulators to hydrodynamic forecasting for lakes and rivers, and (2) to lay the foundations of a simple emulator prototype for establishing a statistical / dynamical relation between two existing models.

ECCC has low-resolution and high-resolution geophysical fields and hydrodynamic simulations, in different domains, including Lake Erie, the St. Lawrence river, Lake Champlain, and the port of St. John (New Brunswick). Model-based computational results will be made available to team members for one of the domains, in order to support their work.

Hydro-Québec (TransÉnergie et Équipement)
Predicting the demand at a given substation

The context

Hydro-Québec TransÉnergie et Équipement (HQTÉ) is trying to predict how a collection of equipments called a medium voltage satellite substation will supply its distribution sector with electricity during the next few hours or days, given the weather conditions and other forecasting parameters. The operation of a substation is characterized by several dependent variables, measured using varied units (power, intensity, etc.). The dependence between variables arises because they are associated to the outputs of one or several equipments that are linked to one another (transformer, circuit breaker, etc.).

The problem

HQTÉ wishes to build a single model that, by learning through measurements, will first simulate the whole set of dependent variables, explained by the independent variables or inputs (temperature, wind, snow cover, precipitations, day of the week, date, hour, profile correction by client type, macroscopic load level). The dependent variables (outputs) are the active power, the reactive power, and the intensity at the exit points.

The required solution

This model will be compared to several models already in use (each of which dedicated to a dependent variable), in order to assess the relevance of the improvements obtained by taking into account the links among the dependent variables. Statistical measures such as the absolute mean of deviations on residuals could be used for this comparison. The results could help partition or sequence the computation of this set of variables. The same procedure could be carried out with forecast data for the inputs in order to measure its reliability in an operational context. Hourly data will be provided by HQTÉ for a two-year horizon.

International Air Transport Association (IATA)
Turbulence in the air: creating a heat map and building a seasonal diagram

This problem actually consists of two subproblems. First IATA wishes the team to create a heat map of turbulence for a given flight level (plus or minus 2k feet). For example, IATA would like to see the heat map of the displayed area for flight level 24 (meaning that all data points from 22k feet to 26k feet are considered). The data set can be provided upon request and will be either the live dataset (4 hours) or historical data (several months). Furthermore IATA would like to see whether the team can use analytic tools (such as Statistics, ML/AI) to spot trends and give advice.

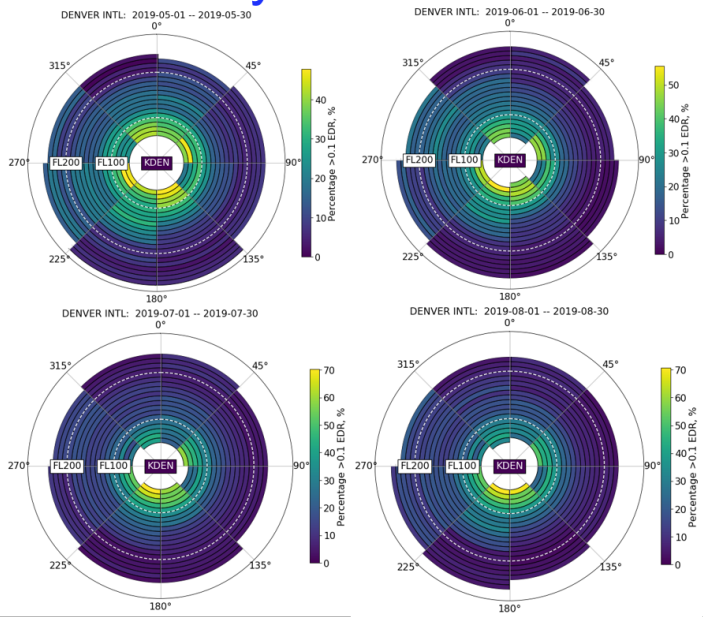
The second part of the problem consists of building a seasonal diagram for a given airport, highlighting the direction of the turbulence, its intensity, and its level. The diagram

below is an example of what IATA is looking for. IATA is interested in the seasonality of turbulence around an airport.

Given a 12-month data set around an airport, the team will have to put together a diagram (or diagrams) showing the characteristics of the turbulence, which we specify now. The turbulence intensity by elevation for each sector (by 2k feet intervals) must be displayed. In the diagram below, each segment is one sector and represents 2k feet; its colour indicates the intensity, which could be either the average or median EDR or the percentage of EDR over a set threshold (e.g., 0.1). The direction of the turbulence around the airport by 45-degrees sectors must be displayed. We need one diagram per season/time frame.

Such diagrams can provide a pilot, when approaching or departing from an airport, with an indication of the expected intensity of the turbulence by flight level and direction. The team should also attempt to provide information on the seasonality of the turbulence for a given airport (for example, we can say that for the Denver airport the likelihood of turbulence during approach is x% from the north in summer). IATA is open to suggestions from the team and wants to tap their expertise in building the most straightforward graph for providing pilots with rich information about turbulence and letting them make decisions without an overload of information.

Aerodrome analysis: Directional Turbulence



Radio-Canada
Automatic text simplification of a public broadcaster's articles

The mandate of CBC / Radio-Canada is to inform, enlighten, and entertain all Canadians. When presenting its plan on equity, diversity, and inclusion for 2022-2025, CBC / Radio-Canada committed itself to doing the utmost so that all persons living in Canada feel valued, recognized, and heard by their public broadcaster from sea to sea. On its web site (radio-canada.ca), Radio-Canada publishes between 450 and 600 articles each day. These articles deal with complex topics (health crisis, climate crisis, the economy, the polarization of society, international conflicts, etc.). Since a good understanding of current issues is necessary to take part in the democratic debate, Radio-Canada thinks that the use of Automatic Text Simplification (ATS) could help a greater number of citizens take part in this debate. Simplifying or summarizing some of its written contents (in an automatic fashion) could enhance the understanding of the articles and make them more attractive for people struggling with literacy, neurodiverse people, and new immigrants (for instance).

In April 2021, CBC / Radio-Canada created Mauril2, a digital platform for learning French and English through audio-visual information produced by the public broadcaster. The development team is currently trying to broaden the supply of written contents through a new reading comprehension task. To this end Radio-Canada will need ATS to produce simpler versions of original articles, in order to take the level of beginners (learning French or English) into account. ATS consists of decreasing the complexity of a text, from the lexical and syntactic points of view, while retaining its meaning. This will improve the readability and ease of understanding of the text. Methods for simplifying text fall into two categories: modular systems (which carry out lexical or syntactic simplifying operations iteratively or recursively), and end-to-end systems, which learn to carry out several modifications at the same time through labeled data. In modular systems the transformations are in most cases applied within sentences.

By taking part in the workshop the Radio-Canada team wishes to explore the most promising avenues for text simplification, especially within the context of a public broadcaster that wants to provide as many citizens as possible with quality information.

Revenu Québec
Identifying companies at risk

Revenu Québec wishes to build a tool for identifying companies at risk (sometimes called "fraudulent companies") based on research already published. The article *Guilt-by-Constellation : Fraud Detection by Suspicious Clique Memberships* ([GbC], see citation below) proposes a method for identifying companies at risk that utilizes a bipartite graph linking groups of companies and collections of resources. Here are the steps of this method: create a graph with vertex weights and edge weights, detect cliques in this graph, evaluate the score of each clique, compute the attributes for each company using the clique information, and use the computed attributes within a predictive model.

The aforementioned article is a sequel of the article *GOTCHA! Network-based Fraud Detection for Social Security Fraud* ([GOTCHA], see citation below), which also proposes a method for identifying companies at risk but does not use the notion of clique. Work based on this article (and almost completed) has been carried out at Revenu Québec and has involved creating a weighted graph, computing most of the attributes, and formulating a model predicting companies at risk.

For the problem proposed for the workshop, one has a particular interest in the attributes based on cliques (see Section 4.4 in [GbC]). Naturally cliques must be identified before these attributes can be computed. In Section 4.3 of [GbC], the authors use a bottom-up approach to find all the cliques in the bipartite graph. This approach has a poor performance and cannot be used on bipartite graphs with tens of thousands of company vertices and millions of resource vertices. Therefore Revenu Québec wishes the team to propose an efficient method for clique detection. The list of clique-based attributes, however, is the desired output: thus Revenu Québec would like to have a global solution that, given a weighted graph, produces such a list for each company.

In order for the team members to test the algorithms they propose, Revenu Québec will provide them with weighted graphs having the same properties as those processed at the department.

[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.