CENTRE FOR MAINTENANCE OPTIMIZATION AND RELIABILITY ENGINEERING

DIRECTOR
CHI-GUHN LEE

SEMI-ANNUAL REPORT
JUNE 2018

DEPARTMENT OF MECHANICAL AND INDUSTRIAL ENGINEERING
UNIVERSITY OF TORONTO
5 KING'S COLLEGE ROAD
TORONTO, ONTARIO, CANADA M5S 3G8
T: +1 (416) 978-2921 F: +1 (416) 946-5462
www.cmore.mie.utoronto.ca

TABLE OF CONTENTS

EXECUTIVE SUMMARY

CHI-GUHN LEE, C-MORE DIRECTOR

INTRODUCTION

The following report summarizes work undertaken between C-MORE and collaborating companies and notes the major changes at C-MORE since the meeting in December 2017.

As Director of the Centre for Maintenance Optimization and Reliability Engineering (C-MORE), I have continued to expand the service and research portfolio of the Centre by ensuring the continuous engagement of consortium members and driving the research program in new directions, such as data analytics and machine learning. Since the last progress meeting, the Consortium has welcomed a new member: Kinross Gold.

At the same time, I have been supervising undergraduate projects, including a machine learning approach to water quality prediction in a refugee camp and statistical methods for flatness control of LINAC, to name a few. With my graduate students, I have been actively involved in reinforcement learning with Bayesian inferencing for multiple information sources, sieve-basis empirical value iteration for dynamic optimization, inverse reinforcement learning, multi-armed bandit for optimal balance between exploration and exploitation, etc.

The C-MORE team has been busy since the December meeting. I am very happy that Professor Fae Azhari, MIE, University of Toronto, Professor Scott Sanner, MIE, University of Toronto, and Professor Sharareh Taghipour, MIE, Ryerson University, continue to be affiliated with C-MORE. I was also pleased to welcome Dr. Janet Lam as Assistant Director of C-MORE in January 2018. Janet knows C-MORE extremely well and is an invaluable addition to the team. Finally, Dragan Banjevic has “officially” retired, but he has willingly continued to work with Consortium members on various projects.

THE C-MORE TEAM

JANET LAM, ASSISTANT DIRECTOR

Dr. Janet Lam began her role as Assistant Director of C-MORE in January 2018. Janet is very familiar with C-MORE; she acquired her doctorate while at C-MORE and then worked as a postdoctoral fellow before taking a teaching position at an American university. Since her arrival at C-MORE, she has focused on developing material for the integration of new consortium members and project progress for existing members. As part of C-MORE’s new efforts in machine learning, she has participated in student supervision as well as self-instruction in the area.

PROFESSOR FAE AZHARI

Fae Azhari is an Assistant Professor, MIE, University of Toronto. Fae is interested in structural health monitoring (SHM) and prognosis of engineering systems. Her main areas of research are twofold: (1) sensor development and assessing the performance of novel sensing devices, and (2)

developing decision-making frameworks that use probabilistic models to translate collected data into meaningful information and efficient remedial strategies for various infrastructure systems. Fae will be conducting a workshop on Structural Health Monitoring at the Progress Meeting.

PROFESSOR SCOTT SANNER

Scott Sanner is an Assistant Professor, MIE, University of Toronto. Scott's research spans a broad range of topics from the data-driven fields of Machine Learning and Information Retrieval to the decision-driven fields of Artificial Intelligence and Operations Research. Scott has applied the analytic and algorithmic tools from these fields to diverse application areas, such as recommender systems, interactive text visualization, and Smart Cities applications, including transport optimization.

PROFESSOR SHARAREH TAGHIPOUR

Sharareh Taghipour is an Associate Professor, MIE, Ryerson University. Her research interests include reliability engineering, inspection and maintenance optimization, stochastic operations research, statistical analysis, and novel applications of maintenance optimization models, such as optimization of cancer screening. She holds a status-only Associate Professor appointment in MIE, University of Toronto, and has collaborated with C-MORE on a number of projects.

DRAGAN BANJEVIC, PROJECT DIRECTOR

Dragan retired from his position at C-MORE at the end of 2017, after 23 years of work. Dragan's input has been invaluable to all students and postdoctoral fellows associated with the program – fortunately for us, his "retirement" has included occupying his old office and working with C-MORE researchers and consortium members.

C-MORE STUDENTS

Michael Gimelfarb, PhD Candidate: Michael is in the PhD program under the co-supervision of Professor Scott Sanner and Chi-Guhn Lee on Bayesian reward shaping for reinforcement learning algorithms, which will be an essential tool for maintenance optimization when the models are unknown. He has successfully completed his research project "Reinforcement Learning with Multiple Experts: A Bayesian Combination Approach" and submitted a paper to the 32nd Annual Conference on Neural Information Processing Systems (NIPS).

Daniel Duklas: Daniel's work at PMH has ended and some of his thesis appears later in the Report; he will also be presenting findings at the meeting. He is graduating with a BASc in Industrial Engineering in June 2018.

Mozam Shahin: Mozam worked with Daniel on the PMH project and has completed the fourth year of his undergraduate program with honours. He will be participating remotely in the presentation at the meeting. He is currently researching at Lund University in Sweden.

MIE undergraduates *Tonglin Jin*, *Yuheng Lin*, *Xuehan Wang*, and *Yuze Li* were part of the CAPSTONE design project with Princess Margaret Hospital (PMH). It was completed in April 2018. Professor Chi-Guhn Lee was supervisor; Professor Daniel Letourneau, also from U of T, represented PMH.

C-MORE ACTIVITIES WITH CONSORTIUM MEMBERS

Since December 2017, C-MORE lab members have been working on research, participating in conferences, and meeting with Consortium members. C-MORE is currently involved in the following projects with industry partners:

BARRICK GOLD

C-MORE received six years of data from Barrick's Veladero Pumping System. There are two pumps with four vibration sensors each. The objective is to analyse the vibration data for a relationship to pump downtime. Though we have a record of pump availability, the cause of downtime, whether it is planned or unplanned is not available. There is one significant period of downtime in the spring of 2017, and several short spurts of the pumps not running. We are focusing on the data leading up to the major outage to see how the two pumps are related, along with temperature and vibration data. Further details will be presented at the progress meeting.

KINROSS GOLD

As a new consortium member, the first half of 2018 was an orientation to C-MORE and our core competencies. We have received a dataset of their Caterpillar haul truck engines and are in the process of preparing it for EXAKT analysis. There are two different types of engines being considered. There are oil and filter changes approximately every 1000 working hours, and preventive maintenance services every 500 hours. This makes the data a great candidate for oil analysis. There are 14 engines, of which there are 8 and 6 of the two types of models. The earliest engines were installed in July 2012 and the longest serving engine has 125,000 working hours to its credit. Details on data analysis and approach will be presented at the progress meeting.

MINISTRY OF DEFENCE UK

We continued the project on Squirrel helicopter flight incidents. A control chart that verifies the random occurrence of dangerous incidents relative to overall number of incidents was produced. Results will be presented and are included in the technical reports. There are several potential new projects for the balance of the year. The Digital Twin project is based on the concept where one has a digital model of a system, so that one can infer condition (and other) information about the original from observing the behaviour of the digital twin. While the common application of the digital twin is at a physical level, we are interested in modelling at the system-level to make good maintenance decisions for a fleet of vehicles. Optimising procurement strategies – there are several strategies under consideration in the area of procurement. Given some input conditions, we are looking at ways to select the strategy most likely to be successful.

TECK

There are several potential projects under consideration for the balance of the year. Economic decision tool – Teck is interested in decision support on whether an asset should be repaired, replaced, or left alone, given a set of inputs. We are also considering maintenance strategies for the replacement of shovel undercarriages. An on-site repair takes 3 weeks to complete, whereas a hot-swap can be done in 10 days. However, the storage and inventory cost of maintaining a spare undercarriage is in question. Remanufacturing strategies are also being considered. It is

uncertain whether the use of remanufactured parts is optimal. It may be preferable to use new parts from different vendors, or that remanufacturing vendors may have differences in quality.

TORONTO TRANSIT COMMISSION

There were two main projects with the TTC this season. One was the completion of the inspection scheduling for the NDT (non-destructive testing) team and the other was the machine learning analysis of infrared video footage and the other. Both projects will be presented at the progress meeting and details can be found in the technical reports.

TORONTO HYDRO

Toronto Hydro's many transformers are governed by external regulations. Many are coming due for replacement, causing a spike in the expected capital investment. Our ongoing project is to develop a plan to smooth the required investment so that the annual budgets for replacements are stable. Details can be found in the technical reports.

C-MORE EDUCATIONAL PROGRAMS

Andrew Jardine has continued to offer courses in asset management around the world. His next Physical Asset Management course at the University of Toronto (in conjunction with the School of Continuing Studies) will run November 5 - 9, 2018. Since 2000, it has been offered as an 8-day program, but this year it has been trimmed to 5 days. Based on ongoing changes in the workplace, the content has been revised; there is less focus on asset management, but continuing emphasis on data analytics, and a new section on machine learning.

EXAKT SOFTWARE

One of the benefits of Consortium membership is access to software. I am pleased to announce an extension of EXAKT. The new version has a number of interesting features; these will be discussed at the Progress Meeting, and more information appears later in the Report.

CONCLUSION

At C-MORE, I have discovered an outstanding team of colleagues, many of whom continue to work with us even after retirement, notably Founding Director Andrew Jardine, but also Dragan Banjevic and Elizabeth Thompson. It is also gratifying that former employees are eager to return, and here I point to Janet Lam. I have enjoyed the past year, working with the collaborating companies and learning about their specific needs. I am confident C-MORE will continue to maintain the support of current industry members through the hard work and dedication of its staff and students.

Chi-Guhn Lee
June 2018

C-MORE LAB ACTIVITIES DECEMBER 2017 – JUNE 2018

VISITS AND INTERACTIONS WITH CONSORTIUM MEMBERS AND OTHERS

Date	Company	Participants	Location	Topic
Jan 08, 2018	TTC	Janet Lam, Miguel Lamsaki, Songchen Liu, Ayaz Rahman, Scott Sanner	Conference call	Machine learning for FLIR videos
Jan 11, 2018	Team Eagle	Dragan Banjevic, Paul Cudmore, Janet Lam, Chi-Guhn Lee, Ty Shattuck	U of T	Preliminary meeting
Jan 19, 2018	CEA	Dragan Banjevic, Daniel Gent, Janet Lam, Chi-Guhn Lee	Conference call	White paper on predictive maintenance for generating units
Jan 25, 2018	TTC	Dragan Banjevic, Janet Lam, Jennifer Lu, Hossein Mohammadian, Aleksandar Urosevic	TTC	Inspection distribution of NDT team
Feb 07, 2018	Kinross	Dragan Banjevic, Arthur Gozzo, Janet Lam, Chi-Guhn Lee, Emilio Sarno, Brian Wright	Kinross	Preliminary meeting
Feb 08, 2018	TTC	Dragan Banjevic, Janet Lam, Jennifer Lu, Hossein Mohammadian	U of T	Inspection distribution of NDT team
Feb 26, 2018	Inha University	Dragan Banjevic, Byunggeun Choi (Gyungsang Univ.), Sun Hur, Sungwoo Kang (Inha Univ.), Janet Lam, Chi-Guhn Lee,Kangmin Lim (ATG)	U of T	Preliminary meeting
Feb 27, 2018	Kinross	Dragan Banjevic, Arthur Gozzo, Janet Lam, Chi-Guhn Lee, Bryan Murphy, Emilio Sarno, Brian Wright	Kinross	EXAKT training
Mar 06, 2018	TTC	Dragan Banjevic, Janet Lam, Jennifer Lu, Hossein Mohammadian, Aleksandar Urosevic	TTC	Inspection distribution of NDT team
Mar 08, 2018	TTC	Aleksandar Urosevic, Robert Capobianco, Tuocheng Liu	TTC	Project detail meeting
Mar 09, 2018	TTC	Robert Capobianco, Janet Lam	TTC	Inspection distribution of NDT team

Mar 12, 2018	Toronto Hydro	Janet Lam, Chi-Guhn Lee, Sakaran Manivannan, Peter Nearing Chris Scarpelli	Toronto Hydro	Potential projects exploration
Mar 13, 2018	Oniqua	Dragan Banjevic, Joe Berti, SiewMun Ha, Andrew Jardine, Janet Lam, Chi-Guhn Lee, Daming Lin	Conference call	Preliminary meeting
Mar 22, 2018	Metrolinx	Alan Britton, Robert Fuller, Xavier Hall, Janet Lam, Chi-Guhn Lee, Martin Sigsworth, Wesley Suh, Duwayne Williams	Metrolinx	Preliminary meeting
Mar 27, 2018	Kinross	Dragan Banjevic, Arthur Gozzo, Hugh Handyside, Janet Lam, Chi-Guhn Lee, Bryan Murphy, Emilio Sarno, Brian Wright	Kinross	SMS training
Mar 28, 2018	Oniqua	Dragan Banjevic, SiewMun Ha, Andrew Jardine, Janet Lam, Chi-Guhn Lee, Daming Lin, James Reyes-Picknell	U of T	Project launch meeting
Mar 29, 2018	Barrick	Thomas A, Dragan Banjevic, Janet Lam	Conference call	Data clarification meeting
Apr 06, 2018	Toronto Hydro	Dragan Banjevic, Janet Lam, Chi-Guhn Lee, Gary Wang	U of T	Project launch meeting
Apr 09, 2018	Metrolinx	Dragan Banjevic, Janet Lam, Alan Marriott, Martin Sigsworth, Alan Swaby	Metrolinx	Project detail meeting
Apr 16, 2018	Metrolinx	Ahmed Ateeq, Jody Fleck, Felix Fung, Janet Lam, Chi-Guhn Lee, Michael McNeill, Arben Muka, Anthony Pezzetti, Derek Polson, Rodrigo Sanos, Bonnie Tam	Metrolinx	Preliminary meeting
Apr 30, 2018	TTC	Dragan Banjevic, Mo Ghaus, David Girodat, Craig Harper, Janet Lam, Jennifer Lu, Hossein Mohammadian, Ayaz Rahman, Aleksadar Urosevic, Horacio Wechow	TTC	Project wrap-up meeting
May 01, 2018	Cerrejon	Dragan Banjevic, Janet Lam, Chi-Guhn Lee, Andres Osorio	U of T	Preliminary meeting

May 04, 2018	TTC	Chi-Guhn Lee, Scott Sanner, Aleksandar Urosevic, Robert Capobianco, Tuocheng Liu	Conference call	Project interim progress meeting
May 08, 2018	CEA	Daniel Gent, Janet Lam, Chi-Guhn Lee	Conference call	Project follow up
May 22, 2018	Teck	Dragan Banjevic, Justin Cvetko Lueger, Graeme Dillon, Kevin Hatch, Will Kearns, Janet Lam, Chi-Guhn Lee, Peter Pistner, Todd Zaccarelli	Conference call	Potential projects exploration
May 25, 2018	Bombardier	Dragan Banjevic, John Coll, Robert Duffield, Philippe Herfray, Kyle Kryway, Janet Lam, Chi-Guhn Lee	Conference call	Preliminary meeting

C-MORE LEADERSHIP ACTIVITIES

CHI-GUHN LEE, C-MORE DIRECTOR

Chi-Guhn Lee, the Director of the C-MORE, has continued to expand the service and research portfolio of the Centre by ensuring the continuous engagement of consortium members and driving the research program in new directions, such as data analytics and machine learning. Since the last progress meeting, the Consortium has welcomed a new member: Kinross Gold.

Chi-Guhn has supervised undergraduate projects, such as a machine learning approach to water quality prediction in a refugee camp and statistical methods for flatness control of LINAC, to name a few. With graduate students, he has been actively involved in reinforcement learning with Bayesian inferencing for multiple information sources, sieve-basis empirical value iteration for dynamic optimization, inverse reinforcement learning, multi-armed bandit for optimal balance between exploration and exploitation, etc.

JANET LAM, ASSISTANT DIRECTOR

Janet returned to C-MORE as the Assistant Director in January 2018. Since her arrival, she has focused on developing material for the integration of new consortium members and on project progress for existing members. As part of C-MORE's new efforts in machine learning, she has participated in student supervision as well as self-instruction in the area.

ANDREW K.S. JARDINE, PROFESSOR EMERITUS

Andrew Jardine, Principal Investigator, Evidence Based Asset Management, and C-MORE's Founding Director, continues to liaise with Director Chi-Guhn Lee and company representatives. His wide network of contacts, gathered over more than 20 years with C-MORE, remains a valuable asset. Andrew has also continued his active role in C-MORE educational programs. He will co-present the extremely popular University of Toronto certificate program in Physical Asset Management, November 5-9, 2018.

DRAGAN BANJEVIC, C-MORE PROJECT DIRECTOR

Despite his official retirement, Dragan has continued to collaborate with members of C-MORE, in particular with Janet. He is mainly contributing to the collaboration with companies.

SHARAREH TAGHIPOUR, RYERSON, EXTERNAL COLLABORATOR

Professor Taghipour is an Associate Professor, MIE, Ryerson University. Her research interests include reliability engineering, inspection and maintenance optimization, stochastic operations research, statistical analysis, and novel applications of maintenance optimization models, such as optimization of cancer screening. She holds a status-only Associate Professor appointment in the Department of Mechanical and Industrial Engineering, University of Toronto, and has collaborated with C-MORE on a number of projects.

SCOTT SANNER, UNIVERSITY OF TORONTO

Scott Sanner is an Assistant Professor, MIE, University of Toronto. Scott's research spans a broad range of topics from the data-driven fields of Machine Learning and Information Retrieval to the decision-driven fields of Artificial Intelligence and Operations Research. Scott has applied the analytic and algorithmic tools from these fields to diverse application areas such as recommender systems, interactive text visualization, and Smart Cities applications including transport optimization.

FAE AZHARI, UNIVERSITY OF TORONTO

Fae Azhari is an Assistant Professor, MIE, University of Toronto. Professor Azhari is interested in structural health monitoring (SHM) and prognosis of engineering systems. Her main areas of research are twofold: (I) sensor development and assessing the performance of novel sensing devices, and (II) developing decision-making frameworks that use probabilistic models to translate collected data into meaningful information and efficient remedial strategies for various infrastructure systems.

MIE490 - CAPSTONE DESIGN PROJECT WITH PRINCESS MARGARET HOSPITAL

The CAPSTONE design project with Princess Margaret Hospital (PMH) was completed in April 2018, with Professor Chi-Guhn Lee as supervisor and Professor Daniel Letourneau, also from U of T, as a representative of PMH. The U of T student team includes Tonglin Jin, Yuheng Lin, Xuehan Wang, and Yuze Li.

PROJECT DESCRIPTION

Radiation therapy uses high energy ionizing radiation (photon and electron beams) to treat patients with cancer. The goal of radiation therapy is to deliver a high dose of radiation to the tumor to eradicate it or control its growth while limiting the delivered dose to surrounding organs that might be sensitive to radiation dose. Medical linear accelerators are the treatment units used to delivery radiation therapy treatments. This complex equipment can deliver lethal doses of radiation to patients if they are miscalibrated. Performances of medical linear accelerators are assessed using a quality control (QC) program with daily, weekly, monthly, and annual tests. Linear accelerators are computer-controlled devices; they record machine parameters and machine states during operation. The objective of this project is to use QC test results and machine-recorded parameters to:

1- Help diagnose the cause (which subsystem) of a change in machine performance and give advice on the appropriate service intervention.
2- Predict timing for servicing intervention in the linear accelerator.

OTHER WORK WITH PMH DATA

In addition to the Capstone project, another team of two students, Mozam S. Shahin and Daniel M. Duklas, completed a statistical analysis of PMH data as a part of their graduation thesis, in coordination with and predating the Capstone team. An abbreviated report of their work appears later.

C-MORE EDUCATIONAL PROGRAMS

Andrew Jardine has continued to offer courses in asset management around the world. In March 2018, he taught a graduate course at the University of the West Indies, Trinidad: MENG 6704: Maintenance Analysis and Optimization. His next Physical Asset Management course at the University of Toronto (in conjunction with the School of Continuing Studies) will run November 5 - 9, 2018. Since 2000, this extremely popular course has been offered as an 8-day program, but this year it has been trimmed to 5 days. Based on ongoing changes in the workplace, the content has been revised; there is less focus on asset management, but continuing emphasis on data analytics, and a new section on machine learning.

NEW FEATURES IN EXAKT V 4.3, APRIL 2018

One of the benefits of Consortium membership is access to software. We are pleased to announce an extension of EXAKT v4.2.1. The main features are the following:

- Windows 7/8/10 compatibility;
- Output RULE and Std Dev to Decisions table;
- Remaining useful life estimate and standard deviation are now saved in two new columns in the Decisions table whenever a Survival Report is generated;
- Locale date issue is resolved; previously, to use EXAKT it was necessary to have the Windows system-wide date format MM/DD/YYYY, and this is not the default in some locales. This is no longer required.
- DMDR database is generated automatically. The first time the ModelDBAttachScript is run, if the script specifies a DMDR db in the current working directory (location of WMOD db), but none is found there, EXAKTm will create an empty one from an Access template file found in the EXAKT installation directory. The intent of this change is to make trying EXAKT on a new project easier;
- Memory leaks fixed;
- Crashes fixed;
- Codebase simplified in a number of ways to enable faster development in the future.

C-MORE GRADUATE STUDENTS

Mike Gimelfarb has successfully completed his research project “Reinforcement Learning with Multiple Experts: A Bayesian Combination Approach” and submitted a paper to the 32nd Annual Conference on Neural Information Processing Systems (NIPS).

OVERALL PROJECT DIRECTION

JANET LAM, C-MORE

GOALS AND RETROSPECTIVES

This report gives a short overview of the activities in C-MORE for the period December 2017 - June 2018. Professor Chi-Guhn continues his leadership activity as the Director of C-MORE. He established several new contacts with industry, which resulted in a new member of the Consortium, **Kinross Gold**, and interest in membership from **Metrolinx**, **DND**, and **Bombardier**. Janet Lam has been appointed a new member of C-MORE staff, primarily oriented towards collaboration with the consortium members. Dragan Banjevic retired in December 2017 but is helping in the transition by acting as a consultant to C-MORE (he was told he has to stay with C-MORE forever). Collaboration with the consortium members has continued on the current projects, and we have discussed ideas for new ones. Research activity has continued with one graduate student while more students will find their place at C-MORE, depending on budgetary opportunities. We are also expecting to hire some postdoctoral researchers to enhance C-MORE's capabilities and growing requirements.

ACTIVITIES

THEORETICAL WORK

This section on theoretical work is oriented towards students' and postdoctoral fellows' research topics and topics of interest for further development.

NAME	ACTIVITY
Michael Gimelfarb, PhD Candidate	Michael Gimelfarb successfully defended his Master's thesis and is in the doctoral program working under the co-supervision of Professor Scott Sanner and Professor Chi-Guhn Lee on Bayesian reward shaping for reinforcement learning algorithms, an essential tool for maintenance optimization when the models are unknown. He submitted a paper on this topic to the NIPS 2018 conference. A more detailed review of his work is included in the report.

INDUSTRY COLLABORATIONS

This section gives details on progress in research conducted with consortium members.

NAME	ACTIVITY
MOD (UK)	The first phase of the project on dangerous flying incidents in Squirrel helicopters fleet has been completed. Dragan has worked on a plausible mathematical model for testing incidence occurrences and the creation of a practical process chart. The data on Squirrel flying hours, received in November, have been analyzed. The results will be presented at the meeting. Tim suggested two new potential projects: the Digital Twin, and Optimizing Procurement Strategies projects.

Teck	Justin Cvetko Lueger proposed several new projects for the upcoming period. Two were selected as top ones, after discussion with C-MORE: Economic Decision Tool, and Maintenance Strategies for the Replacement of Shovel Undercarriages. Some details on other options are included in the report.
Toronto Hydro	A project on smooth replacement of aging transformers fleet has been initiated to avoid big spikes in capital investment. The project will be presented at the meeting.
TTC	The project on optimal frequency and schedule of subway track inspections has been completed. Further analysis of the NDT MOWIS data is of interest, such as failure modes, time transitions between defects etc. Machine learning analysis of infrared video footage has been conducted. The results of both projects will be presented at the meeting.
Barrick	C-MORE received 6 years of data from Barrick's Veladero Pumping System of two pumps; these include vibration and temperature data. The data on pump availability over time are included, but the cause of downtime, whether planned or unplanned, is not available. Further details will be presented at the meeting.
Kinross Gold	C-MORE introduced Kinross Gold to the potential benefits of collaboration. As a result, we have received a dataset of their Caterpillar haul truck engines and are in the process of preparing it for EXAKT analysis. Details on data analysis and our approach will be presented at the meeting.

LAB GOALS FOR SUMMER AND FALL 2018

The theoretical research and collaborations with companies will continue, depending on C-MORE staff availability. The software development options will be explored, for example, developing a prototype incidents process control chart for MOD UK, following the analysis completed in the previous period. This chart will be useful to other companies in similar settings.

C-MORE ACTIVITIES WITH COLLABORATING COMPANIES: AN OVERVIEW

JANET LAM, C-MORE

These are the key projects with consortium members since January 2018.

BARRICK GOLD

C-MORE received six years of data from Barrick's Veladero Pumping System. There are two pumps with four vibration sensors each. The objective is to analyse the vibration data for a relationship to pump downtime. Though we have a record of pump availability, the cause of downtime, whether it is planned or unplanned is not available.

There is one significant period of downtime in the spring of 2017, and several short spurts of the pumps not running. We are focusing on the data leading up to the major outage to see how the two pumps are related, along with temperature and vibration data. Further details will be presented at the progress meeting.

KINROSS GOLD

As a new consortium member, the first half of 2018 was an orientation to C-MORE and our core competencies. We have received a dataset of their Caterpillar haul truck engines and are in the process of preparing it for EXAKT analysis.

There are two different types of engines being considered. There are oil and filter changes approximately every 1000 working hours, and preventive maintenance services every 500 hours. This makes the data a great candidate for oil analysis. There are 14 engines, of which there are 8 and 6 of the two types of models.

The earliest engines were installed in July 2012 and the longest serving engine has 125,000 working hours to its credit. Details on data analysis and approach will be presented at the progress meeting.

MINISTRY OF DEFENCE UK

We continued the project on Squirrel helicopter flight incidents. A control chart that verifies the random occurrence of dangerous incidents relative to overall number of incidents was produced. Results will be presented and are included in the technical reports.

There are several potential new projects for the balance of the year.

The Digital Twin project is based on the concept where one has a digital model of a system, so that one can infer condition (and other) information about the original from observing the behaviour of the digital twin. While the common application of the digital twin is at a physical

level, we are interested in modelling at the system-level to make good maintenance decisions for a fleet of vehicles.

Optimising procurement strategies – there are several strategies under consideration in the area of procurement. Given some input conditions, we are looking at ways to select the strategy most likely to be successful.

TECK

There are several potential projects under consideration for the balance of the year.
Economic decision tool – Teck is interested in decision support on whether an asset should be repaired, replaced, or left alone, given a set of inputs.

We are also considering maintenance strategies for the replacement of shovel undercarriages. An on-site repair takes 3 weeks to complete, whereas a hot-swap can be done in 10 days. However, the storage and inventory cost of maintaining a spare undercarriage is in question.

Remanufacturing strategies are also being considered. It is uncertain whether the use of remanufactured parts is optimal. It may be preferable to use new parts from different vendors, or that remanufacturing vendors may have differences in quality.

TORONTO TRANSIT COMMISSION

There were two main projects with the TTC this season. One was the completion of the inspection scheduling for the NDT (non-destructive testing) team and the other was the machine learning analysis of infrared video footage and the other. Both projects will be presented at the progress meeting and details can be found in the technical reports.

TORONTO HYDRO

Toronto Hydro’s many transformers are governed by external regulations. Many are coming due for replacement, causing a spike in the expected capital investment. Our ongoing project is to develop a plan to smooth the required investment so that the annual budgets for replacements are stable. Details can be found in the technical reports.

TECHNICAL REPORTS: CONSORTIUM

ANALYSIS OF SQUIRREL INCIDENTS WITH RESPECT TO FLYING HOURS

DRAGAN BANJEVIC, C-MORE
JANET LAM, C-MORE

BACKGROUND

Tim Jefferis from MOD UK sent the Squirrel monthly flying hours at the end of 2017 to help analyse the assumption of a Poisson distribution of incidents in flying time. The data provide the fleet cumulative flying hours per month from April 2012 to October 2017. The incidence data cover the period from July 2011 to July 2017, slightly different from the flying hour period. In this analysis, the flying hours are used for the period covered by incidence data. The table below gives the summary for every month of

1. flying days,
2. flying hours (original data),
3. count of "regular" incidents ("near-collisions" excluded) – as found in flying incidence data,
4. rate per day (= count/days),
5. expected number of "regular" incidents (calculated as the product of flying days and the overall (average) rate per day (calculated for the entire period, see below),
6. standardized value of the counts (calculated as (count – expected)/sqrt(expected), see below).

Rate of incidents per day is calculated for every month separately, as, e.g., for April 2012, 14/36.743 = 0.38102, and overall rate is calculated for entire interval as total count/total hours = 1453/2983.84 = 0.48695 (see the end of the table).

The expected number of incidents per month is calculated from assumptions that all months counts follow Poisson distribution with rate equal to the overall rate, e.g., for April 2012, it is 36.743 x 0.48695 = 17.892. [all calculations are done with higher precision and given rounded in the table]

The difference between actual count and expected value can be used to check our Poisson assumption, as well as possible deviation from typical/expected number of regular incidents. But this deviation also depends on the expected number of incidents – clearly, with a larger value, the deviation can be larger. To adjust for this, we calculate the "standardized" deviation as (Count – expected)/sqrt(expected); e.g., for April 2012, it is (14 -17.892)/sqrt(17.892) = -0.9202. Division by the square root of expected comes from our assumption of a Poisson distribution. (For Poisson distribution with rate parameter λ , expected value is λ, and standard deviation is sqrt(λ)). So the standardized counts are a good method to look at deviations; see below.

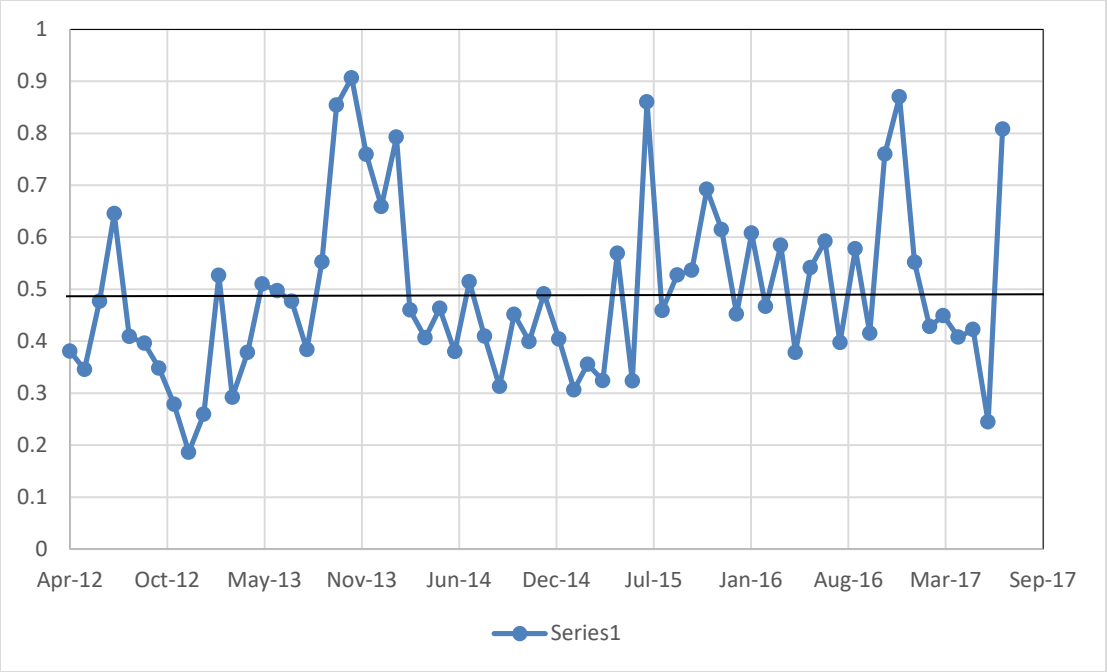
For simplicity and clarity, we present the analysis in a graphical form.

Month	Days	Hours	Count_reg	rate/day	expected/month	standardized
Apr-12	36.743	881.8333	14	0.38102	17.892	-0.9202
May-12	52.108	1250.583	18	0.34544	25.374	-1.4639
Jun-12	41.927	1006.25	20	0.47702	20.417	-0.0922
Jul-12	49.594	1190.25	32	0.64524	24.15	1.59739
Aug-12	53.771	1290.5	22	0.40914	26.184	-0.8177
Sep-12	45.441	1090.583	18	0.39612	22.128	-0.8775
Oct-12	43.052	1033.25	15	0.34842	20.964	-1.3027
Nov-12	43.094	1034.25	12	0.27846	20.985	-1.9613
Dec-12	21.448	514.75	4	0.1865	10.444	-1.994
Jan-13	26.948	646.75	7	0.25976	13.122	-1.6901
Feb-13	47.469	1139.25	25	0.52666	23.115	0.39203
Mar-13	34.212	821.0833	10	0.2923	16.66	-1.6316
Apr-13	39.622	950.9167	15	0.37858	19.294	-0.9776
May-13	45.094	1082.25	23	0.51005	21.959	0.22222
Jun-13	52.302	1255.25	26	0.49711	25.469	0.10525
Jul-13	48.243	1157.833	23	0.47675	23.492	-0.1016
Aug-13	59.944	1438.667	23	0.38369	29.19	-1.1458
Sep-13	47.042	1129	26	0.5527	22.907	0.64619
Oct-13	44.497	1067.917	38	0.854	21.668	3.50861
Nov-13	46.316	1111.583	42	0.90681	22.554	4.09471
Dec-13	25.024	600.5833	19	0.75926	12.186	1.95206
Jan-14	40.962	983.0833	27	0.65915	19.947	1.5793
Feb-14	47.948	1150.75	38	0.79253	23.349	3.03215
Mar-14	47.813	1147.5	22	0.46013	23.283	-0.2658
Apr-14	44.281	1062.75	18	0.40649	21.563	-0.7673
May-14	41.014	984.34	19	0.46325	19.972	-0.2175
Jun-14	60.507	1452.167	23	0.38012	29.464	-1.1909
Jul-14	60.26	1446.25	31	0.51443	29.344	0.30567
Aug-14	41.503	996.0833	17	0.4096	20.21	-0.7141
Sep-14	44.736	1073.667	14	0.31295	21.785	-1.6679
Oct-14	50.965	1223.167	23	0.45129	24.818	-0.3649
Nov-14	37.566	901.5833	15	0.3993	18.293	-0.7699
Dec-14	32.573	781.75	16	0.49121	15.862	0.03475
Jan-15	44.566	1069.583	18	0.4039	21.702	-0.7946
Feb-15	55.458	1331	17	0.30654	27.006	-1.9254
Mar-15	47.813	1147.5	17	0.35556	23.283	-1.302
Apr-15	55.507	1332.167	18	0.32428	27.029	-1.7368
May-15	33.396	801.5	19	0.56893	16.262	0.67888

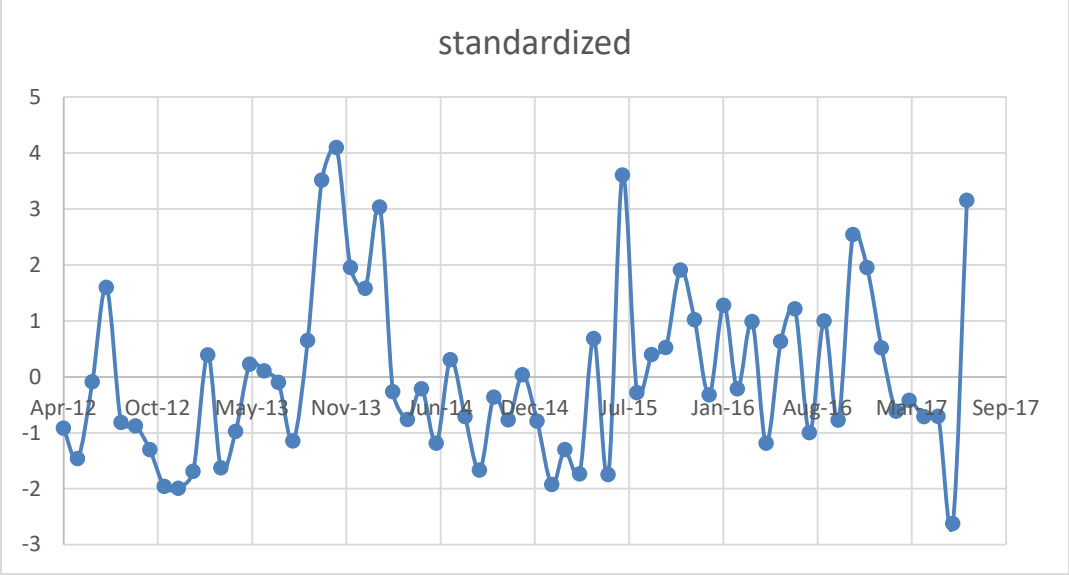
Jun-15	55.632	1335.167	18	0.32356	27.09	-1.7465
Jul-15	45.333	1088	39	0.86029	22.075	3.60218
Aug-15	50.128	1203.083	23	0.45882	24.41	-0.2855
Sep-15	47.427	1138.25	25	0.52712	23.095	0.39642
Oct-15	54.073	1297.75	29	0.53631	26.331	0.5201
Nov-15	41.896	1005.5	29	0.69219	20.401	1.90369
Dec-15	30.903	741.6667	19	0.61483	15.048	1.01868
Jan-16	42.031	1008.75	19	0.45204	20.467	-0.3243
Feb-16	54.285	1302.833	33	0.60791	26.434	1.27702
Mar-16	55.681	1336.333	26	0.46695	27.114	-0.2139
Apr-16	49.628	1191.083	29	0.58434	24.167	0.98314
May-16	58.181	1396.333	22	0.37813	28.331	-1.1895
Jun-16	64.628	1551.083	35	0.54156	31.471	0.62902
Jul-16	64.132	1539.167	38	0.59253	31.229	1.21155
Aug-16	60.396	1449.5	24	0.39738	29.41	-0.9976
Sep-16	58.875	1413	34	0.57749	28.67	0.99553
Oct-16	57.764	1386.333	24	0.41548	28.128	-0.7784
Nov-16	42.118	1010.833	32	0.75977	20.51	2.5372
Dec-16	12.639	303.3333	11	0.87033	6.1546	1.95313
Jan-17	30.816	739.5833	17	0.55166	15.006	0.51474
Feb-17	53.736	1289.667	23	0.42802	26.167	-0.6191
Mar-17	60.111	1442.667	27	0.44917	29.271	-0.4198
Apr-17	39.25	942	16	0.40764	19.113	-0.7121
May-17	59.17	1420.083	25	0.42251	28.813	-0.7104
Jun-17	57.205	1372.917	14	0.24473	27.856	-2.6253
Jul-17	47.045	1129.083	38	0.80773	22.909	3.15295
Total	2983.84		1453	0.486956315		

ANALYSIS

The first graph below presents daily incidence rates over months. If all is “normal”, the rates are expected to be randomly “scattered” around the overall rate of 0.4869/day (the middle line). The graph shows that there are intervals with rates below average and above average, which may be expected with random variations. There are several months with either very low or high rates. Are they also expected?



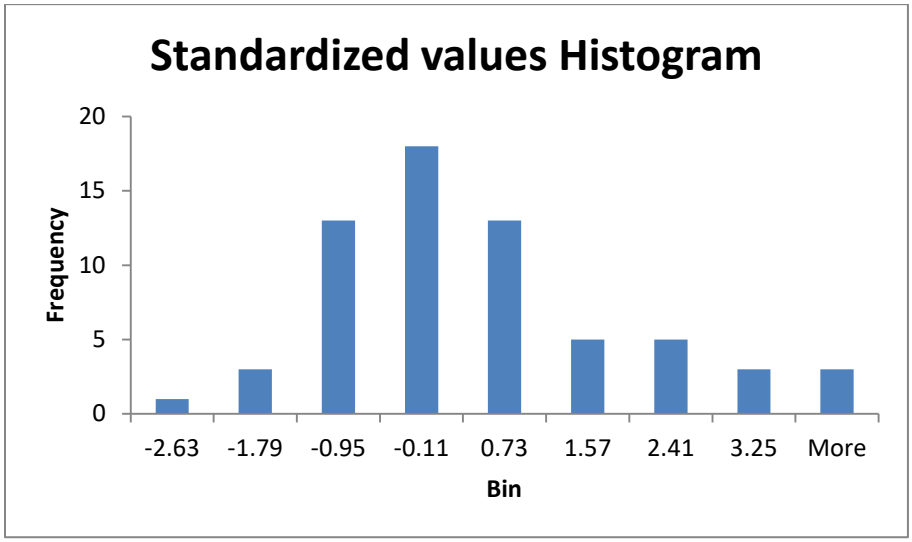
The second graph (shown below) provides standardized counts. As mentioned above, the standardized counts give better description of deviations.



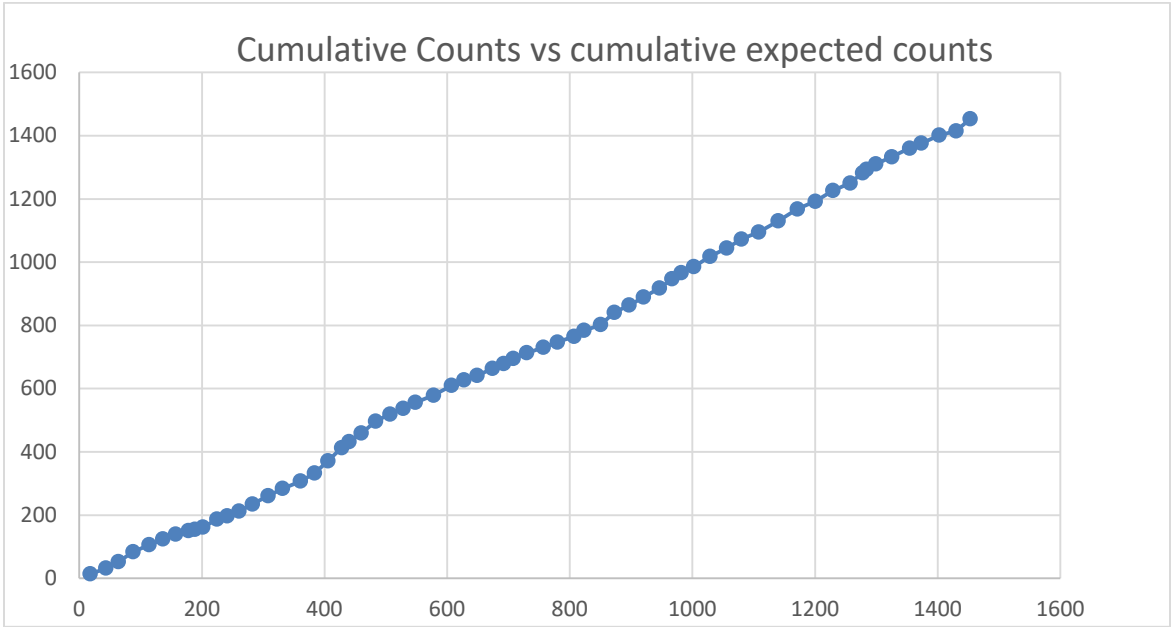
Statistical theory says standardized counts should follow standard Normal distribution (mean = 0, and standard deviation = 1), approximately, when the expected count is not very small, and when we have a Poisson distribution. For normal distribution, the interval of [-2, +2] (approximately) should contain 95% of the cases. Using a common statistical reasoning, we might suspect that the standardized counts greater than 2 (or smaller than -2) are "un-normal" – or suspicious – and there are several of them in our case. As the most look randomly scattered around zero value, we may conclude that overall, the incidents follow Poisson distribution, with a couple of exceptions where the number of incidents is rather high or low. See, e.g., October

and November 2013, July 2015, and July 2017 for high values, and June 2017 for a low value. As the number of points (months) in the analysis is 64, we would expect not more than 5 points outside the limits. Here we have 7, 2 more than the upper limit, supporting our claim of some months with un-normalities in the number of incidents.

The third graph (shown below) is the histogram of the standardized values. The distribution looks normal, with exception of the slightly longer right tail, due to a couple of exceptional months with high values, as discussed above.



The final graph that can be useful for our analysis shows the cumulative monthly counts vs the cumulative expected monthly counts, as described above (cumulatives given in Appendix). If our assumption about the Poisson distribution is correct, the data points should follow a straight line with random variation.



As can be seen from the first and second graphs (above) and the Appendix, the actual monthly cumulative counts of incidents lag behind (less than expected) in the period; the lag is up to 400 expected incidents (November 2013). They later catch up and show steady behaviour (relatively small deviation from the straight line).

CONCLUSION

The analysis shows that the assumption of the Poisson process for the "regular" incidents is reasonable and can be used in further development of the process chart, as indicated at the December 2017 C-MORE meeting. The analysis also shows possibilities for creating various process charts, if flying hours are available.

APPENDIX

CUMULATIVE COUNTS OF INCIDENTS

Month	Count_reg	expected/month	Cum_exp	Cum_count
Apr-12	14	17.89226	17.89226	14
May-12	18	25.37414	43.26641	32
Jun-12	20	20.41666	63.68306	52
Jul-12	32	24.14999	87.83305	84
Aug-12	22	26.18405	114.0171	106
Sep-12	18	22.12777	136.1449	124
Oct-12	15	20.96448	157.1094	139
Nov-12	12	20.98477	178.0941	151
Dec-12	4	10.4442	188.5383	155
Jan-13	7	13.12246	201.6608	162
Feb-13	25	23.11521	224.776	187
Mar-13	10	16.65965	241.4356	197
Apr-13	15	19.29395	260.7296	212
May-13	23	21.95869	282.6883	235
Jun-13	26	25.46883	308.1571	261
Jul-13	23	23.49226	331.6494	284
Aug-13	23	29.19033	360.8397	307
Sep-13	26	22.90724	383.7469	333
Oct-13	38	21.66787	405.4148	371
Nov-13	42	22.55386	427.9687	413
Dec-13	19	12.18574	440.1544	432
Jan-14	27	19.94661	460.101	459
Feb-14	38	23.34854	483.4496	497
Mar-14	22	23.2826	506.7322	519
Apr-14	18	21.56303	528.2952	537

May-14	19	19.97211	548.2673	556
Jun-14	23	29.46424	577.7315	579
Jul-14	31	29.34419	607.0757	610
Aug-14	17	20.21038	627.2861	627
Sep-14	14	21.78453	649.0706	641
Oct-14	23	24.81786	673.8885	664
Nov-14	15	18.29299	692.1815	679
Dec-14	16	15.86159	708.0431	695
Jan-15	18	21.70168	729.7448	713
Feb-15	17	27.00579	756.7505	730
Mar-15	17	23.2826	780.0331	747
Apr-15	18	27.02946	807.0626	765
May-15	19	16.26231	823.3249	784
Jun-15	18	27.09033	850.4152	802
Jul-15	39	22.07535	872.4906	841
Aug-15	23	24.41038	896.901	864
Sep-15	25	23.09492	919.9959	889
Oct-15	29	26.33115	946.327	918
Nov-15	29	20.40144	966.7285	947
Dec-15	19	15.0483	981.7768	966
Jan-16	19	20.46738	1002.244	985
Feb-16	33	26.43429	1028.678	1018
Mar-16	26	27.114	1055.792	1044
Apr-16	29	24.1669	1079.959	1073
May-16	22	28.33139	1108.291	1095
Jun-16	35	31.47124	1139.762	1130
Jul-16	38	31.22946	1170.991	1168
Aug-16	24	29.41013	1200.402	1192
Sep-16	34	28.66955	1229.071	1226
Oct-16	24	28.12849	1257.2	1250
Nov-16	32	20.50965	1277.709	1282
Dec-16	11	6.154587	1283.864	1293
Jan-17	17	15.00603	1298.87	1310
Feb-17	23	26.16714	1325.037	1333
Mar-17	27	29.27149	1354.308	1360
Apr-17	16	19.11304	1373.422	1376
May-17	25	28.81327	1402.235	1401
Jun-17	14	27.85627	1430.091	1415
Jul-17	38	22.90893	1453	1453

MOD PROJECT: CREATING A PROCESS CHART FOR TESTING THE FREQUENCY OF AN EVENT WITH RESPECT TO ANOTHER EVENT

DRAGAN BANJEVIC, C-MORE
JANET LAM, C-MORE

BACKGROUND

Two (or more) types of events (accidents) are recorded in the flying time of aircraft (Squirrel helicopters). For the first type of event (type 1), we may expect relatively steady and random appearance in flying time, but for the second type (type 2) we may suspect some non-random variations, clustering, external causes, etc. Under "normal" conditions, type 2 events should also appear with random fluctuations, so we are interested in checking possible deviations from random appearance, preferably in the form of a process control chart. The data for actual testing were provided by Tim Jefferis from DSTL/MOD UK in August 2017. Tim provided limited data on flying hours at the end of 2017 for testing assumptions of a Poisson process.

STATEMENT OF THE PROBLEM AND METHODOLOGY

An approach to the solution of the problem was reported at December 2017 C-MORE Meeting. In this report we will take a slightly different approach, using binomial distribution, and suggest a simple method of presenting accidents data on a process chart. The key assumption is that events of both types occur at random in cumulative flying time *t*, and they occur independently. A common reasonable assumption for this situation is that these processes are homogeneous Poisson processes (HPP) with appropriate occurrence rates. The analysis of incidents with known flying hours has shown that the Poisson assumption is reasonable; the analysis is included in this progress report.

The number of events of type 1 in time *t* is X_t, with occurrence rate λ and of type 2, Y_t, with occurrence rate μ. If time *t* is available, we can check every, say, 100 hours whether Y_t, the number of events of type 2 in 100 hours, is compatible with the assumed Poisson process with rate μ, or not. But as emphasized by Tim, the operating time *t* is hard to obtain or could be unreliable, so we cannot check Y_t directly. Tim's idea was to use (more common) events of type 1 as *a time variable*, that is, to compare occurrence of type 2 events with occurrence of type 1 events. This methodology was discussed in the December 2017 report. In mathematical terms, it means if it is known at a certain moment in calendar time that $X_t = k$, what is the distribution of Y_t, or probability of $Y_t = i$, for *i*=0, 1, 2, …, *but without knowledge of t*? After discussion at the December meeting, Tim suggested conditioning on the total number of incidents of both types as a more common method. That is, if it is known that $Z_t = X_t + Y_t = n$, what is the distribution of Y_t, or probability that $Y_t = i$? In this form, the solution is simple, and this probability does not depend on *t*, as in the previous approach:

$$P(Y_t = i | Z_t = n) = b(i; p, n) = \binom{n}{i} p^i q^{n-i}, i = 0,1,2, \dots, n \ ,$$

the binomial distribution with parameters *n* and *p*; *p* is the probability that if an event occurs, it is of type 1, and *q* is the probability it is of type 2. We will see in the Appendix that $p = \frac{\mu}{\lambda+\mu}$ and $= \frac{\lambda}{\lambda+\mu}$. In that case,

$E(Y_t|Z_t = n) = np,\ Var(Y_t|Z_t = n) = npq, Stdev(Y_t|Z_t = n) = \sqrt{npq}$.

In plain language, if all is "normal" (no exceptional rise or drop in incidents of type 2), the resulting numbers of incidents should follow a binomial distribution. We expect Y_t will be around its expected value of *np*, more or less. We can use a 90% or 95% confidence interval.

ESTIMATION OF PARAMETERS

The parameter *p* is either estimated from a previous study or set as default value. Assume that in the observation period, type 1 event occurred n_1 times and type 2 event n_2 times. The obvious (we will not give formal arguments) estimates are

$\hat{p} = \frac{n_1}{n_1+n_2}$, $\hat{q} = \frac{n_2}{n_1+n_2}$,

where we may exclude the cases (time intervals) in which our assumption about "normal" behavior is clearly violated. This means there is a reasonable doubt that type 2 events did not occur at random.

Example: The Squirrel data report 1581 incidents of both types, with 33 incidents of type 2, and 1548 incidents of type 1 in a calendar time span from 18/07/2011 to 31/07/2017, almost exactly 6 years. The probability p of the event of type 2 is estimated as $= \frac{33}{1581} = 2.0873\%$, if we don't exclude any time interval. We can check the incidence data in fixed calendar intervals.

TESTING METHODOLOGY

At the given checking date, e.g., every 3 months (Tim's suggestion), the number of events of type 1, *k*, and type 2, *i*, is observed, and *n* = *k* + *i*. Then the probability of not exceeding the observed value *i* is calculated, given the observed value *n*, or

$$P(Y_t \le i|Z_t = n) = \sum_{j\le i} \binom{n}{j} p^j q^{n-j} .$$

If the probability $P(Y_t \le i|Z_t = n)$ is large, this means $P(Y_t \ge i|Z_t = n)$ is small, say, less than 0.05, and we may have an unusual behaviour, i.e., too many incidents of type 2. If the probability $P(Y_t \le i|Z_t = n)$ is smaller than the selected critical value, e.g., 0.05, we may raise a red flag and investigate type 2 incidents for possible unusual behaviour. As Tim noticed, a small value of *i* can be suspicious, e.g., due to a lack of reporting. So we may use a one-sided or a two sided method for testing. In two-sided testing, with critical value 0.05, we check whether $(Y_t \le i|Z_t = n) \le 0.025$, or $P(Y_t \ge i|Z_t = n) \le 0.025$.

Fine-tuning of this method is possible by checking type 2 events, e.g., every 3 months, 6 months, and 1-year intervals, per single interval and cumulatively. This method would help to avoid being alarmed by random fluctuations that might look suspicious in a short time span but are actually okay over a longer time span.

APPLYING THE TESTING METHODOLOGY TO THE SQUIRREL DATA SET

From Squirrel data, we have estimated (see above) $p = \frac{33}{1581} = 0.020873$. Starting with the first record in the data, on 18/07/2011, and using checking points at 3-month intervals (approximately), in January, April, July, and October every year, and ending on 31/07/2017, we obtain the results shown in Table 1.

The table shows the counts of all events and the count of events of type 2 in columns 2 and 3, and the appropriate cumulative probabilities of occurrence of type 2 events, given the observed count of all events, in column 4. The probabilities outside a 95% confidence interval (either below 2.5% or above 97.5%) are highlighted. If we look at all checking points, we see those on 03/10/11 and 07/01/2013 are above the 97.5% limit, while one on 02/07/2015 is just below it. This might mean there were some unusual type 2 events in these intervals. Even so, having 2 "unusual" intervals in 25 intervals is not all that unusual (if all is regular, in our testing an "unusual" interval would have a probability of 5%) and has the cumulative probability of 0.87289; having 3 "unusual" intervals out of 25 is 0.9659, still below the 97.5% limit, but closer. If we look at yearly intervals, only the first one (01/07/2011 - 02/07/2012) has a higher cumulative probability value of 91.3% (6 type 2 events out of 181 events in the previous year), still inside the 90% confidence interval (between 5% and 95%). Overall, in this data set we cannot find anything overly irregular.

Date	Total events	Type 2 events	Cumulative probability	Cumulative prob., 1 year
03/10/2011	47	3	0.983400	
04/01/2012	34	1	0.841917	
05/04/2012	48	1	0.735067	
02/07/2012	52	1	0.704063	0.913322
01/10/2012	72	1	0.555101	
07/01/2013	31	3	0.996193	
03/04/2013	42	0	0.412327	
02/07/2013	64	1	0.612931	0.727376
01/10/2013	72	2	0.809468	
07/01/2014	99	2	0.658526	
01/04/2014	87	1	0.455568	
02/07/2014	60	2	0.869724	0.653117
01/10/2014	62	0	0.27041	
05/01/2015	54	0	0.320118	
01/04/2015	52	0	0.333912	
02/07/2015	55	3	0.972124	0.314088

06/10/2015	87	1	0.455568	
07/01/2016	77	2	0.782583	
01/04/2016	78	1	0.51379	
01/07/2016	86	1	0.461806	0.31817
03/10/2016	96	2	0.675648	
04/01/2017	67	0	0.243343	
03/04/2017	67	2	0.835417	
03/07/2017	55	2	0.89246	0.614625
31/07/2017	37	1	0.819593	

Table 1: Application of testing methodology to Squirrel data set

PROCESS CHART DEVELOPMENT

The methodology can be simply implemented as a prototype in Excel to create a process chart that can automatized the plotting and calculation. The Excel file has two spread sheets, the first called “Control chart” and intended for testing. The second one, “plotting values”, includes some technical data for plotting and should not be altered. Instructions for using the control chart are included and the probability *p* calculated, as explained above in estimation section. In this version of the prototype, the probability *p* is fixed, but it can be made flexible for possible updates. The control chart includes data for 3-month intervals and the appropriate cumulative probabilities as explained in the previous section; the last interval covers only one month, up to the end of reported incidents.

The user can enter the next checking interval date (it need not be 3 months but should not be very short), the total number of events, and the number of events of type 2 in that interval. The cumulative probability will be calculated automatically; the point on the graph will also appear. The cumulative probability scale (the “y” scale) is slightly distorted to more clearly show small and large probabilities, as they are the main interest.

The prototype can be improved to include adjustment of probability, *p*, to include the simultaneous checking of other intervals, such as yearly. We hope to get feed-back from MOD to make the prototype practical and useful for other consortium members.

APPENDIX

Derivation of the binomial distribution for conditional probabilities:

Let X_t follow the Poisson distribution (λ) , and Y_t follow the Poisson distribution $P(\mu)$; that is $P(X_t = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$, and $P(Y_t = i) = \frac{(\mu t)^i}{i!} e^{-\mu t}$. Then, $Z_t = X_t + Y_t$ also follows the Poisson distribution $P(\lambda + \mu)$, or $P(Z_t = n) = \frac{((\lambda+\mu)t)^n}{n!} e^{-(\lambda+\mu)t}$.

Then,

$$P(Y_t = i|Z_t = n) = \frac{P(Y_t = i, Z_t = n)}{P(Z_t = n)} = \frac{P(Y_t = i, X_t = n - i)}{P(Z_t = n)} =$$

$$= \frac{P(Y_t = i)P(X_t = n - i)}{P(Z_t = n)} = \frac{\frac{(\mu t)^i}{i!} e^{-\mu t} \frac{(\lambda t)^{n-i}}{(n-i)!} e^{-\lambda t}}{\frac{((\lambda+\mu)t)^n}{n!} e^{-(\lambda+\mu)t}} = \frac{n!}{i!\,(n-i)!} \frac{(\mu t)^i (\lambda t)^{n-i}}{((\lambda+\mu)t)^n}$$

$$= \binom{n}{i} \frac{(\mu)^i (\lambda)^{n-i}}{(\lambda+\mu)^n} = \binom{n}{i} \left(\frac{\mu}{\lambda+\mu}\right)^i \left(\frac{\lambda}{\lambda+\mu}\right)^{n-i} = \binom{n}{i} p^i q^{n-i} = b(i; p, n), p = \frac{\mu}{\lambda+\mu} \ .$$

TTC: AN APPROACH TO INSPECTION SCHEDULE OPTIMIZATION

DRAGAN BANJEVIC, C-MORE
JANET LAM, C-MORE

INITIAL ASSUMPTIONS

We propose a basic methodology to optimize an inspection schedule based on certain simple assumptions about the subway railway system and its faults. The approach can be expanded and improved if more elements of the system's state and operation are considered. A practical solution can be obtained after completing statistical analysis, and calculating the appropriate defect frequencies. The solution is general in nature and can be applied to any system with separate inspection units and limited inspection resources; here we use the terminology of a subway system for clarity.

1. The entire subway rail system (SRS) is divided into non-overlapping segments (e.g., between subway stations).
2. When inspected, the whole segment is inspected in one visit (i.e., not just part of the segment).
3. Frequency of defects per year (defect rate) is known for every segment. In a more advanced study, different types of defects and their rates are be considered. Here we look at all HP (high priority) defects.
4. Defect rate is assumed constant for a given section, under given conditions (such as track type, track geometry, usage and age) and may change when conditions change (e.g., age).
5. "Cost" of inspecting a segment is known (roughly proportional to its length) and is given in men hours, or man days.
6. Total amount of men hours/days available for all inspection over one year is fixed. Or the total "cost" of inspecting SRS in men hours/days over one year is fixed. We call this the budget.

FURTHER ASSUMPTIONS AND OPTIMIZATION CRITERIA

Here we consider more technical assumptions and introduce notation to formulate and solve the problem mathematically.

1. Defects come randomly in time and cannot be predicted.
2. Inspection frequency does not affect defect rate but decreases time between onset of the defect and its discovery. We call this the "unsupervised defect time" or dormant time (DT).
3. The current inspection schedule of inspecting the entire SRS once a year makes 6 months of DT for every defect, on average (1/2 of inspection time unit). If a section of SRS is inspected k times a year, a "uniform" schedule, the DT for its defects is $1/(2k)$ of one year, on average. For example, if a segment is inspected 2 times in one year, the DT of every defect will be 3 months, on average.

4. SRS has N segments, with defect rates λ_i , $i=1,2,...,N$. The defect rate can be defined as #defects/length unit/year. We will also look at "unstandardized" rate = #defects/year = $\lambda_i L_i = \gamma_i$; this definition is less convenient if we want to compare different segments.
5. We look at defects of different types, with rates per year γ_{ij}, j being a defect type, such as "red", "yellow", and "purple", at segment i. We assign a "weight" R_j to each defect type – importance in comparison with other types, such as 10 to red, 5 to yellow and 1 to purple. Then, the total "weighted" defect rate per year for segment i is $B_i = \gamma_{i1}R_1 + \gamma_{i2}R_2 + \cdots$. The weights can be defined differently for different segments (i.e., R_{i1}, R_{i2}, ... for segment i).
6. Let the total amount of men hours/days available for inspection of SRS (subject to inspections) in one year be C, and the "cost" in men hours/days of inspecting segment i be C_i. Then, the cost of inspecting SRS once a year is $\sum_1^N C_i$.
7. Assume now that we inspect different segments with possibly different frequencies. Let the segment i be inspected $k_i, k_i > 0,$ times a year, $i=1,2,...,N$. If $k_i > 1$, the segment is inspected more than once a year, e.g., twice if $k_i = 2$. If $k_i < 1$, the segment is inspected less than once a year; e.g., for $k_i = 0.5$, the segment is inspected *once in two years*. For $k_i = 0.666.. = 2/3$, the segment is inspected 2 times in 3 years, or once every 18 months.
8. The key amount for our inspection optimization is the total DT time for the system, for given schedule of inspection frequencies, $k_1, k_2, ..., k_N$, which is

$DT_{SRS} = \sum_1^N \frac{B_i}{2k_i} = \frac{1}{2}\sum_1^N \frac{B_i}{k_i}$,

following 3 above. Our goal is to find the optimal schedule $k_1^*, k_2^* ..., k_N^*$ from $\min_{k_1,...,k_N} DT_{SYS}$, under the constraint $\sum_1^N k_i C_i = C$.

OPTIMIZING INSPECTION SCHEDULE

Using the objective function of total dormant time, or total unsupervised defect time, DT_{SRS}, described above, we get the following result:

1. The optimal inspection frequencies, k_i^* , for different segments are equal to

   $k_i^* = C * D_i = C * \frac{1}{\sum_1^N \sqrt{B_i C_i}} * \sqrt{\frac{B_i}{C_i}}$.

   The minimal total dormant time is (from B8.)

   $$DT_{SRS}^* = \sum_1^N \frac{B_i}{2k_i^*} = \frac{1}{2C}\left(\sum_1^N \sqrt{B_i C_i}\right)^2.$$

   The result can be easily obtained using optimization under constraints for multidimensional functions (will be provided).
2. For the "uniform" inspection schedule, from $k_i = k$, and $k\sum_1^N C_i = C$, or $k = \frac{C}{\sum_1^N C_i}$,

   $DT_{SRS} = \sum_1^N \frac{B_i}{2k} = \frac{1}{2C}\sum_1^N B_i \times \sum_1^N C_i.$
3. The ratio

$$EFF = \frac{DT_{SRS} - DT^*_{SRS}}{DT_{SRS}} = 1 - \frac{DT^*_{SRS}}{DT_{SRS}} = 1 - \frac{\left(\sum_1^N \sqrt{B_i C_i}\right)^2}{\sum_1^N B_i \times \sum_1^N C_i}$$

measures the efficiency of the optimal schedule over the "uniform" schedule, for the same yearly budget. It is important to note that the efficiency *does not depend* on actual budget size *C*, but only on one-time inspection costs and defect rates.

APPENDIX

MATHEMATICAL DERIVATION OF THE OPTIMAL INSPECTION SCHEDULE

It can be easily done by using the method of "Lagrangian multipliers" for constrained optimization of multidimensional functions. In our case, we want to minimize the function

$$DT_{SRS} = G(k_1, k_2, \cdots, k_N) = \sum_1^N \frac{B_i}{2k_i} = \frac{1}{2}\sum_1^N \frac{B_i}{k_i},$$

with a constraint $\sum_1^N k_i C_i = C$, where also $k_1, k_2, ..., k_N > 0$. With one constraint (one equation), we introduce one dummy variable α, and then minimize the extended function $F(k_1, k_2, \cdots, k_N, \alpha) = \frac{1}{2}\sum_1^N \frac{B_i}{k_i} + \alpha(\sum_1^N k_i C_i - C)$, without constraints. Using partial derivatives, $\frac{\partial}{\partial k_i} F(k_1, k_2, \cdots, k_N, \alpha) = \frac{1}{2}\left(-\frac{B_i}{k_i^2} + \alpha C_i\right) = 0$, we come to the solutions $k_i^* = \frac{1}{\sqrt{\alpha}}\sqrt{\frac{B_i}{C_i}}$, for $i = 1,2, \ldots, N$, depending on dummy value $\sqrt{\alpha}$. After replacing k_i^* into the equation $\sum_1^N k_i C_i = C$, we get $\sqrt{\alpha} = \frac{1}{C}\sum_1^N \sqrt{B_i C_i}$; hence, the final solution

$$k_i^* = \frac{C}{\sum_1^N \sqrt{B_i C_i}}\sqrt{\frac{B_i}{C_i}},$$

for k_i^*. The minimal cost is easily obtained by replacing k_i^* into function $G(k_1, k_2, ..., k_N)$ above, that is,

$$G(k_1^*, k_2^*, \cdots, k_N^*) = \sum_1^N \frac{B_i}{2k_i^*} = \frac{1}{2C}\left(\sum_1^N \sqrt{B_i C_i}\right)^2.$$

A formal argument that it is indeed the minimum (not a maximum) is simple, but we will not include it here.

TTC: DETECTING POWER RAIL ANOMALIES FROM FLIR THERMAL IMAGES USING TENSORFLOW OBJECT DETECTION API

TUOCHENG LIU, MENG STUDENT

BACKGROUND

TTC Line 3 Scarborough is a 6.4KM light metro line connecting six stations from Kennedy to McCowan. It's about 40 years old now and will be replaced by the planned extension of Line 2 Bloor-Danforth subway to Scarborough Centre. While the new project is being built, TTC must keep Line 3 operational.

In May 2017, a train was damaged during operation because of a melted rail caused by a power rail anomaly. This resulted in passenger evacuation and line closure.

PROBLEM

TTC works proactively to eliminate service interruptions on Line 3. One of the methods is scanning the railway periodically, using a FLIR infrared camera installed at the tail of the train. In doing so, power rail anomalies can be identified from thermal images and corrected by maintenance technicians mitigating any unexpected closures.

However, manual observation on these infrared videos requires long-span focused attention, which makes this task more suitable for computer vision applications. The objective of this project is to develop an anomaly detector which works on railway images recorded by an FLIR infrared camera. By feeding in the infrared video, the detector should be able to identify the appearance and location of the anomaly if it exists in a specific frame.

APPROACH

In recent years, many research papers on computer vision, especially object detectors, have been published, and the source code for GitHub repositories is available for public use. TensorFlow Object Detection API is an open source framework presented by Google which covers the process of constructing, training and evaluating object detection models. This codebase also features a variety of modern convolutional models; these allow developers to train new object detectors with their own dataset, on top of the structures of pre-trained models provided in the application program interface (API). This is also the approach for developing the anomaly detector in this project.

For data collection, TTC provided one pair of infrared videos along Line 3, both northbound and southbound, recorded in December 2017. Each video has about 16,000 frames, and 45 anomalies in total were identified by manual observation. To construct a dataset, all frames were exported, and those frames with anomalies were selected for labelling. The labelling process was done using Microsoft Visual Object Tagging Tool (VoTT).

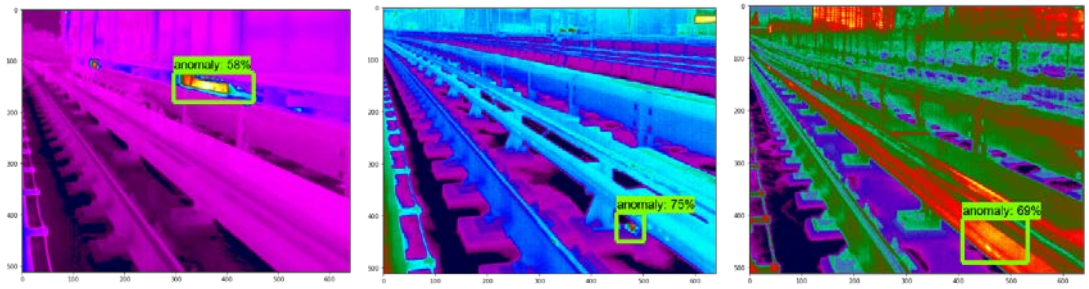
As of the end of May, 780 frames were selected for labeling based on clarity of the anomaly in the image. A few prototype models have been trained to determine if there are enough data to train an effective detector and to fine-tune the hyper parameters in the training process, such as learning rate, batch size, initial ratios of bounding boxes and loss function coefficients.

RESULTS

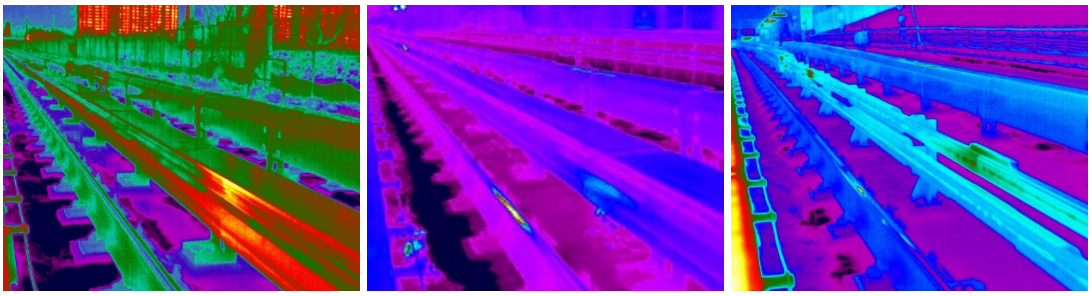
To evaluate the performance of the prototype anomaly detector, a small dataset was constructed by selecting the five best quality frames for each anomaly. Then the 45 anomalies were randomly divided into four folds, such that each fold could be used as a test set once and the remaining three folds could be used as training sets. The following table reports the true positive and false positive of the models trained in four iterations. Note that false positive is measured using 200 randomly selected frames without anomaly.

Cross Validation	True positive	False positive
Iteration 1	81.8%	2%
Iteration 2	90.8%	2.5%
Iteration 3	75.5%	6%
Iteration 4	73.9%	3%

Examples of false positives:



Examples of false negatives:



DISCUSSION AND FUTURE STEPS

The above evaluation shows promising results for the anomaly detector to be delivered in August. The differences between iterations show that larger datasets will help the model to generalize better.

After investigating the false positive examples, three sources of error can be categorized: background objects, tiny hotspots and ambiguity. For false negatives, the model also tends to miss the ambiguous ones which might be hard to determine even for humans.

In the next steps, more models will be iteratively trained and evaluated on 10 and more frames for each anomaly to explore the best performance achievable with the data given. Then, additional filter and ranking rules will be established based on observations to reduce errors in the detection results. If time permits, a simple GUI will be developed so that operators at TTC can use the model easily. The model evaluation in the final report is expected to be based on another set of infrared videos to be provided.

TTC NON-DESTRUCTIVE TESTING INSPECTION PROJECT

JANET LAM, C-MORE
DRAGAN BANJEVIC, C-MORE

BACKGROUND

At the TTC, the non-destructive testing (NDT) team's time is somewhat of a scarce resource; they currently spend a fixed amount of time performing line-tests of the complete subway rail system once per year. By adjusting the inspection frequency of each part of the rail system, we may be able to achieve better reliability with the same amount of resources.

This report summarizes all of the steps that were taken to perform this analysis and its results.

STEPS OF ANALYSIS

The main steps taken to perform this analysis are as follows:

1. Edit asset register so that every point of the track appears exactly once
2. Define scope of analysis - what parts of the system is included in the project?
3. Filter MOWIS defects to those within scope of analysis - which entries on the MOWIS databased are included in the project?
4. Match each defect to an inspection section
5. Prepare matched spreadsheet for mathematical analysis - minor adjustments for anomalies
6. Perform mathematical analysis using number of defects in each inspection segment

This report will discuss each of these steps in turn.

ASSET REGISTER EDITING

The asset register that C-MORE received followed a protocol that didn't quite meet our needs. The objectives of the asset register were as follows:

- A continuous sequence of track in both bounds for the lines in question. That is, each ending chainage must match the beginning chainage of the next entry in the sequence
- Characteristics measured – track geometry, structure type, chainages, direction, reference to stations
- A new row (or entry) whenever a characteristic changed

Some specific modifications we made are as follows:

- If a curved track enters into a station, it was split into two entries – the section that is adjacent to the station, and the section that is part of the station. Both entries would have the same track geometry.
- Entries of alternating tangents and curves (for example see Yorkdale to Lawrence West) were separated into individual entries, with chainages that follow sequentially.

The whole asset register was edited following a similar philosophy.

RAIL SYSTEM TRACK SELECTION

For the purposes of this project, only a selection of the entire system was analysed. This section summarizes how the scope of analysis was derived.

- The analysis was limited to Line 2 (Bloor-Danforth), Line 3 (Scarborough SRT), Line 4 (Sheppard) and Line 1 (Yonge-University-Spadina) from Sheppard West station (previously Downsview station) to Finch station. The newest extension from Sheppard West to Vaughan Metropolitan Centre was not considered
- Exclusion of special tracks - The NDT inspection team provided a schedule describing the time required to inspect each section of the system. This schedule separated the time to perform mainline inspections from special tracks. Only the time for mainline inspections and tails were included. Specifically, crossovers, yards, storage, bridges were excluded from our analysis. Tails were included because the MOWIS defect list does not distinguish tails from the mainline.
- The total number of days used for NDT mainline inspections was counted to be 92.

FILTERING DEFECTS FROM MOWIS DATABASE

In order to match the defects to the available inspection time, we needed to filter out the defect entries to those strictly for analysis. In this section, we summarize which defects were considered, and which were excluded.

- Defects from January 1 2015 to Feb 7 2018 were included. This is 1134 days.
- Defects occurring in the Toronto-York Spadina Subway Extension (TYSSE) were excluded
- Only NDT defects were included. Entries arising from projects were excluded.
- Only defects located in the MAINLINE zone were included.
- Only defects with clearly defined bounds NB, SB, EB, WB were included. Centre tracks or cross overs with UNK bounds were excluded.
- For each defect history, only the initial entry (with status NEW) was counted. In the same spirit, only the initial location markers & offsets were considered, the updated markers were not used.
- The result was 1235 entries.

MATCHING DEFECTS TO RAIL SEGMENTS

Using programming, each of the 1235 defects were matched to a location on the asset register. The asset register kept count of every defect that occurred along the rail system, along with the priority colour.

The input files were prepared as follows:

- NDTnoProj.csv - MOWIS excerpt with formatting removed. Received directly from TTC, excludes projects.
- assetsForMOWIS.csv - Subset of asset register with formatting removed. Line, track and chainage information only.
- stationAssets.csv - Asset register with station labels.
- stationDaystoInspect.csv - Asset register with required inspection days for each section.

There is one output file: defectsCounted.csv. It produces a list of all sections with the total number of defects and a count of defects for each priority level.

The code for matching process is provided via three R files. The R files and their overall purposes are as follows:

- initialize.R – loads libraries, data files, selects important columns and removes less important columns
- helpers.R – a file of functions used in the main project file. Each function is described within the file.
- ndtProj.R – the main project file. Runs the first two R files listed above, matches defects to rail segments, then aggregates them to station-to-station segments.

The detailed documentation is included within the R files.

OUTPUT EDITING

Once ndtProj.R had been run and the output file defectsCounted.csv produced, a couple of details were edited to adjust for the system's distinctive characteristics.

Defects that occurred along Union station and St. George station were counted separately from the track that is adjacent to those station. The number of defects occurring in Union station were split in half and shared equally in the counts of defects from St. Andrew--Union and Union--King. A similar approach is advised for St. George; there were no defects recorded in this station.

MATHEMATICAL ANALYSIS

Using the counts of defects per inspection section (station-to-station), an optimal distribution of inspection frequencies was computed. The details of the mathematics are provided in a separate document. In this section, we will provide an overview of the computations.

- For each inspection section, the number of defects for each priority level was counted.

- The priority levels were given a weight for importance, giving a weighted defect count for each inspection section.

Red	Yellow	Purple	Blue	Brown	Gray
10	5	1	0.5	0.25	0.125

Using these weights, a weighted defect count for each section was computed. The method for computation is illustrated in the table below:

Priority	Weight	Defect count	Weighted count
Red	10	1	$10 \times 1 = 10$
Yellow	5	2	$5 \times 2 = 10$
Purple	1	7	$1 \times 7 = 7$
Blue	0.5	10	$0.5 \times 10 = 5$
Brown	0.25	8	$0.25 \times 8 = 4$
Gray	0.125	8	$0.125 \times 8 = 1$
Total weighted defect count			**37**

- For each inspection section, the number of days required to inspect was recorded.
- The relative inspection frequency for the *i*-th inspection section was computed by

$$f_i = \frac{1}{\sum_i \sqrt{d_i w_i}} \sqrt{\frac{d_i}{w_i}}$$

where d_i is the number of days required to inspect the *i*-th section, and w_i is the weighted number of defects for the *i*-th section.

- f_i can be multiplied by D, the number of days available per year for mainline inspections, to compute n_i, the total number of inspections per year. In this analysis, $D = 92$.

$$n_i = f_i \times D$$

- The optimal inspection interval I_i in days is

$$I_i = \frac{1}{n_i} \times 365$$

The objective to be minimized in this project is unsupervised time, or the time between a defect occurrence and its detection by the NDT team.

On a policy in which each section is inspected equally, the unsupervised time U_e in our planning horizon of 1134 days is calculated by

$$U_e = \frac{\sum_i d_i \sum_i w_i}{2D} = 349$$

The unsupervised time for the optimal inspection policy is

$$U^* = \frac{\left(\sum_i \sqrt{d_i w_i}\right)^2}{2D} = 261.6$$

- The resulting percent improvement is

$$E = \left|\frac{U_e - U^*}{U_e}\right| \times 100 = 25.1\%$$

INDEPENDENT ANALYSIS

Given the tools provided in this document and the R files, some independent analysis can be performed by TTC. The relevant values can be changed in the ndtProj.R file to explore how various inputs will change the results. Here, we summarize some of the recommended adjustments.

- Rounding the optimal inspection frequencies to some other reasonable values, such as the nearest week or month. This will have minor consequences to the objective value.
- Adjusting the weighting values of priorities. To ignore some priorities, set to zero.
- Adjusting the total number of days available per year to perform mainline NDT tests.
- Include yards, crossovers, etc. This adjustment is more involved; it requires selecting different rows of the MOWIS file and a change to the available inspection days.

OPTIMAL INSPECTION FREQUENCIES IN DECREASING ORDER

Section	Line	Optimal frequency (days)
Donlands to Greenwood	BD	143.1
Union to King	YUS	164.1
North York Centre to Finch	YUS	165.6
Dundas West to Lansdowne	BD	170.7
Eglinton to Lawrence	YUS	180.7
Sherbourne to Castlefrank	BD	181
St. Clair to Davisville	YUS	182.5
Lawrence to York Mills	YUS	204.8
York Mills to Sheppard	YUS	206.6
Sheppard to North York Centre	YUS	208.8
Victoria Park to Warden	BD	227.1
Wilson to Yorkdale	SPADINA	227.6
Castlefrank to Broadview	BD	233.7
St. Andrew to Union	UNIV	239.1
Davisville to Eglinton	YONGE	240.2
Broadview to Chester	BD	242.6
Chester to Pape	BD	245
Coxwell to Woodbine	BD	246.2
St. George to Bay	BD	258.1
Woodbine to Main	BD	271.9
Wellesley to Bloor	YONGE	278.7
Lansdowne to Dufferin	BD	278.7
Summerhill to St. Clair	YONGE	286.2
Osgoode to St. Andrew	UNIV	290.1
Dufferin to Ossington	BD	290.1
Christie to Bathurst	BD	315.2
Glencairn to Eglinton West	SPADINA	317.8
Spadina to St. George	SPADINA	317.8
Keele to Dundas West	BD	320.5
Yorkdale to Lawrence West	SPADINA	329
Warden to Kennedy	BD	336.6
Finch Tail	YONGE	341.4
Old Mill to Jane	BD	341.4
Eglinton West to St. Clair West	SPADINA	341.4
St. George to Museum	UNIV	355.3
Bloor to Rosedale	YONGE	355.3
St. Patrick to Osgoode	UNIV	359.1
Bay to Yonge	BD	359.1

Section	Line	Optimal frequency (days)
Yonge to Sherbourne	BD	363
Bathurst to Spadina	BD	384.5
Main to Victoria Park	BD	384.5
Scarborough Centre to McCowan	SRT	389.2
Kennedy to Lawrence East	SRT	399.4
Lawrence West to Glencairn	SPADINA	399.4
Pape to Donlands	BD	404.7
Ossington to Christie	BD	422.2
Downsview to Wilson	SPADINA	425.3
College to Wellesley	YONGE	428.5
Kipling to Islington	BD	428.5
Greenwood to Coxwell	BD	449.5
Dupont to Spadina	SPADINA	465.2
Spadina to St. George	BD	492.4
King to Dundas	YONGE	502.5
Rosedale to Summerhilll	YONGE	502.5
High Park to Keele	BD	580.2
Midland to Scarborough Centre	SRT	597.1
Royal York to Old Mill	BD	657.9
Islington to Royal York	BD	682.8
Yonge to Bayview	SHEPPARD	696.3
Museum to Queen's Park	UNIV	742.3
Runnymede to High Park	BD	742.3
Jane to Runnymede	BD	870.4
Leslie to Don Mills	SHEPPARD	965.6
Lawrence East to Ellesmere	SRT	1049.7
Ellesmere to Midland	SRT	1100.9
Downsview Tail	SPADINA	1100.9
Dundas to College	YONGE	1100.9
St. Clair West to Dupont	SPADINA	1230.9
Kipling Tail	BD	1230.9
Queen's Park to St. Patrick	UNIV	1421.3
Kennedy Tail	BD	1421.3
Bayview to Bessarion	SHEPPARD	1740.7
Sheppard Tail	SHEPPARD	No defects
Bessarion to Leslie	SHEPPARD	No defects
Don Mills Tail	SHEPPARD	No defects
McCowan Tail	SRT	No defects

TWO NEW IDEAS FOR A C-MORE / MOD UK COLLABORATIVE PROJECT

TIM JEFFERIS, DSTL/MOD

BACKGROUND

This short report describes two possible collaborative projects between C-MORE and DSTL/MOD.

DIGITAL TWIN RESEARCH

An innovation viewed with increasing interest in the defence field is the concept of a Digital Twin. This is a virtual representation of physical assets (the Physical Twin) that can be used to provide improved information to management decision-making, as it can be interrogated and analysed more conveniently and cheaply than the physical system. The digital twin can be used to support virtual testing and qualification of physical assets and to support their in-service management.

It appears that the majority of current applications of this concept are bottom-up, in that they rely on a detailed understanding of the physics and mechanics of the asset and are used to predict failures and their impacts, diagnose and prognose system condition (based on limited real-time data) and to schedule maintenance actions, for example, through damage accumulation methods.

Whilst this approach is valid and very useful, in relevant circumstances, it appears likely that a different approach would be required to build a useable Digital Twin of a fleet of trucks. If, for example, one wishes to improve the decisions made about the operation and maintenance of such a fleet, it is possible that a top-down Digital Twin might provide sufficient detail and accuracy, so that the detailed work of building a full digital representation of each truck would not be required.

If this approach is pursued, research is required to determine what information management will require the Digital Twin to produce and how accurate this needs to be. It is possible that this could be captured in a Management Critical Information Requirements (MCIR) document. Once it has been established how the Digital Twin needs to behave, then it will be necessary to determine the minimum requirements for data gathered from the Physical Twin, and how timely these data need to be.

PROCUREMENT STRATEGY PROBLEM

A number of procurement strategies can be adopted to deliver a very technologically and/or operationally complex end product. The following three represent the main options currently in vogue:

1. Waterfall Model – In this approach, the requirements for the final product are set; design, production and introduction into service activities then follow sequentially, delivering the final product in a single iteration.
2. Incremental Model – In this approach, the requirements for the final product are set, together with the requirements for an interim level of performance to be delivered. The design work for the initial level of performance is undertaken, informed by the requirements for the final level of performance, followed by the production and introduction into service of the initial capability. Once this has been achieved, the design, production and introduction into service of the final version can be undertaken.
3. Iterative Model – This approach is composed of a series of waterfalls that iteratively work towards a final solution, but rather than setting the final requirements as the first activity, each iteration has a new set of requirements, informed by the results of the previous iterations.

A range of internal and external factors and uncertainties can affect the efficiency and appropriateness of each potential approach in a given situation.

An optimisation process is required to examine the risks and uncertainties in a given situation and calculate the range of potential outcomes for each approach to indicate the range of potential approaches and determine which risks must be especially considered in each case.

KINROSS: CATERPILLAR HAUL TRUCK ENGINES – DATA PREPARATION

JANET LAM, C-MORE

BACKGROUND

In April 2018, C-MORE began a new project with Kinross on the optimal replacement frequency of a fleet of haul truck engines. Kinross supplied event histories and records of inspections gathered at oil changes for its fleet of haul trucks. The data are well-suited for EXAKT analysis to condition-based maintenance using oil analysis data.

The project is currently at the data preparation and analysis stage, including cleanup, identification of anomalies and potential errors. In this report, we summarize the findings in the initial data analysis.

DATA DESCRIPTION

Overall data description

- There are two different engine models: 793C and 793D. The expected ranges of metal particles for these two models are quite different and may benefit from separate analysis.
- There are 15 units – 9 engines are model 793C and 6 are 793D. Three units are retired before 2018.
- There are 34 histories of which 8 ended in failure, 13 ended in suspension and 13 were still in service when the data were reported.
- There are 1323 inspections recorded.
- The data range from July 2012 to February 2018.
- There are 37 columns of measurement variables. Some are more complete than others.
- There are two types of failures in the events table: failures and internal failures. Both were treated as the same type of failure.
- There are three types of suspensions: high-hour, retired and test engine returned. All are treated as the same type of suspension.

Notable data characteristics

Upon closer inspection of the data, some interesting features are found.

- From the beginning of the data in July 2012 until the beginning of 2016, the average inspection interval is about 11 days. Then from 2016, inspection intervals increase to about 25 days. (See Figure 1)
- There are no immediate concerns with multiple simultaneous failures (these may indicate some external issue). In other words, there is no obvious correlation with failures and calendar dates.

- All of the units begin their history with at least 20,000 hours of working age. These may be true hours, meaning the previous data records do not exist, or simply where the odometers started.
- Some measurements, like iron, follow distributions that are highly skewed to the right, and others, such as zinc follow normal distributions (Zn). Some variables appear to have outliers that might suggest a possible measurement error. Some variables such as nickel, silver, cadmium, vanadium and antimony show only a few values such as 0, 1, and 2. (See Figure 2)
- The usage of the haul trucks (operating hours per time unit) appears to remain steady throughout the duration of the data.

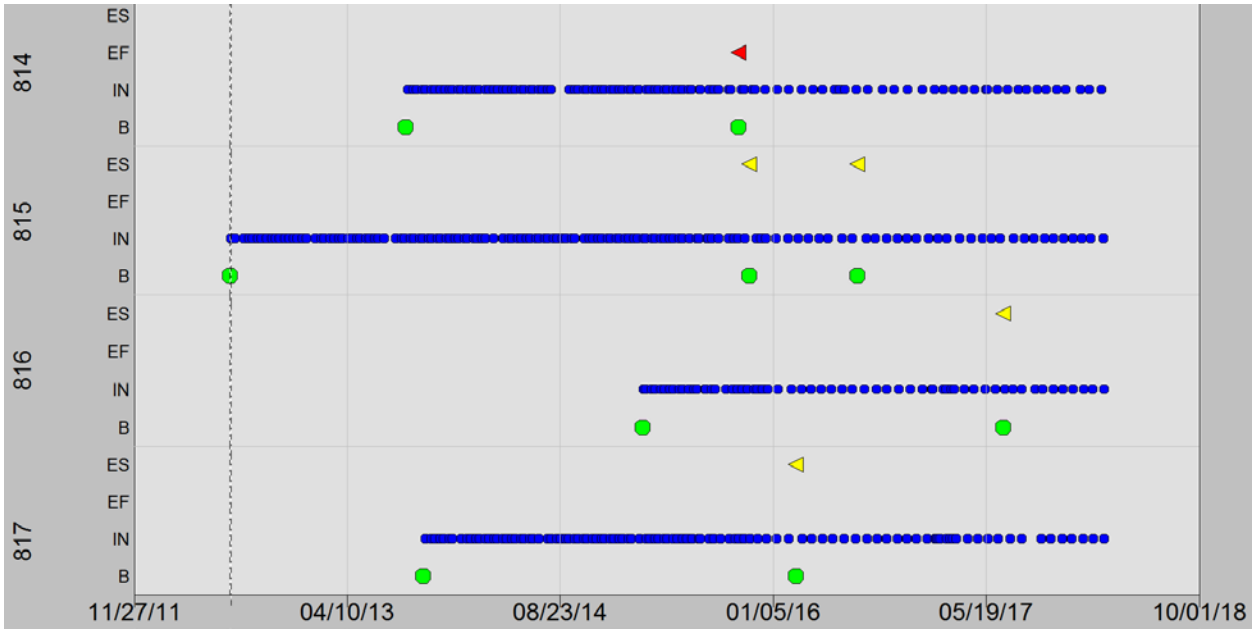


Figure 1 Snapshot of four units, demonstrating change in inspection frequency

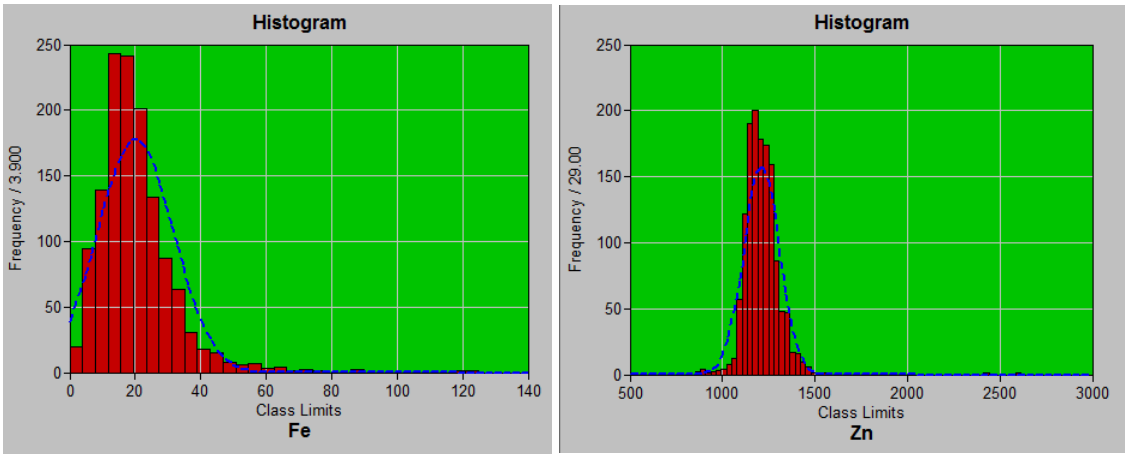


Figure 2 Sample frequency distributions of measurement values

DATA INTEGRITY ISSUES

Each history has a beginning event, an ending event with optional inspections in between. Each event, including inspections, is recorded with a calendar date and working age. There are a few problems that can occur that may indicate some errors in the data.

In this project, there were four main types of issues:

- Ending events followed by inspections, with no beginning events;
- Multiple inspection records reported at the same time, but with different results;
- Inconsistent, or non-sequential working ages with calendar times;
- Working age does not increase between dates.

Each concern is addressed in detail below.

Inspections with no beginning events

Each unit has at least one beginning and an end (whether in suspension or failure), but those that were not retired have inspections that follow the final ending event, without any beginning event. This suggests that there are perhaps missing B and EF/ES events in the events table. It could be assumed that the units were re-installed some time between an ending event and the next inspection. However, if there are missing EF/ES events, it's possible that some histories between the last recorded ending event and the last recorded inspection are unaccounted for.

Based on feedback from Kinross, a new beginning event was entered for each of these cases. Table 1 lists all of these cases.

Table 1 List of cases with ending events followed by inspections with no beginning event

Unit #	Cause of ending event	Date of final ending event	First recorded inspection following ending event	Final inspection reported.
804	High hour	Sep 8, 2016	Sep 29, 2016	Feb 12, 2018
805	High hour	May 25, 2016	June 24, 2016	Dec 8, 2016
807	Failure	Nov 4, 2016	Nov 17, 2016	Feb 21, 2018
809	High hour	Oct 12, 2015	Oct 30, 2015	Feb 22, 2018
810	Internal failure	Sep 16, 2017	Oct 3, 2017	Jan 30, 2018
811	High hour	Aug 12, 2016	Sep 2, 2016	Feb 22, 2018
812	Failure	Aug 19, 2016	Aug 21, 2016	Feb 19, 2018
813	Internal failure	June 15, 2017	June 17, 2017	Feb 26, 2018
814	Failure	Oct 15, 2015	Oct 21, 2015	Feb 13, 2018
815	Test engine return	Jul 20, 2016	Aug 9, 2016	Feb 17, 2018
816	High hour	Jun 28, 2017	Jun 30, 2017	Feb 18, 2018
817	High hour	Feb 26, 2016	Mar 12, 2016	Feb 19, 2018

Multiple inspection records at the same time

There are two occurrences of two inspections being recorded at the same time, but with different values. Perhaps one was incomplete/incorrect measurement; alternatively, the date/time/working age information may be incorrect.

Based on feedback from Kinross, one of each record was selected arbitrarily. There was no reasonable way to determine the nature of the conflict, and it is only two records in 1300 inspections.

Table 2 Sample of measurement values for inspections with two records

			Selection of measurements											
Unit #	**Date**	**Working age**	**Fe**	**Cu**	**Pb**	**Sn**	**Cr**	**Ni**	**Ti**	**Al**	**Si**	**Na**	**K**	**B**
807	May 11, 2015	107453	4	0	0	0	0	0	0	1	3	2	1	52
807	May 11, 2015	107453	33	2	1	1	1	1	0	4	12	5	2	44
814	Apr 10, 2015	58752	9	1	0	1	0	0	0	1	7	6	1	47
814	Apr 10, 2015	58752	14	2	2	0	0	0	0	3	2	6	2	43

Inconsistent/non-sequential working ages

In some cases, working ages would decrease as calendar dates increased.

In one case, there is an obvious typo, but in other cases it's not as clear. In most cases, the inconsistency is within the inspections table. This is identified by two consecutive inspection records with working ages moving backwards in time.

In other cases, the inconsistency is between the inspections table and the events table. This is when an event occurs on the events table, and the following inspection occurs next in dates, but backwards in time.

Table 3 lists pairs of records to show decreasing working age.

Based on Kinross' feedback, working age anomalies were corrected by interpolation.

Table 3 Table of working age anomalies

Unit #	Event	Date	Working age	Comments
805	IN	Jun 24, 2016	115762	
	IN	Jul 18, 2016	497	Most definitely a typo
807	IN	Nov 26, 2014	105002	
	IN	Dec 10, 2014	106074	
807	EF	Nov 4, 2016	118172	
	IN	Nov 17, 2016	117979	
809	ES	Oct 12, 2015	108819	
	IN	Oct 30, 2015	101069	
809	IN	Dec 16, 2015	111791	All surrounding entries are in the 100 thousands
	IN	Dec 18, 2015	102168	
810	IN	Dec 21, 2017	117049	
	IN	Jan 9, 2018	116869	
811	IN	Jul 23, 2015	98550	
	IN	Aug 1, 2015	98249	
812	IN	May 28, 2014	93537	
	IN	Jun 3, 2014	92836	
812	IN	Jan 14, 2015	96509	
	IN	Jan 21, 2015	96504	
812	IN	Feb 5, 2016	105429	
	IN	Mar 7, 2016	104563	
812	IN	May 9, 2017	112839	Also WA for April 17 2017 is 112354, highly improbable
	IN	Jun 8, 2017	112355	
814	IN	Mar 19, 2014	52654	
	IN	Mar 28, 2014	52249	
815	IN	Jun 28, 2012	42332	
	B	July 9, 2012	41838	An inspection before an installation?
815	IN	Apr 16, 2015	59387	The three working ages are all lesser than WA on Apr 16, 2015
	IN	Apr 23, 2015	51498	
	IN	May 7, 2015	51778	
	IN	May 14, 2015	51910	
815	IN	Oct 29, 2015	62713	
	ES	Nov 9, 2015	62648	
817	IN	Jul 16, 2016	48994	
	IN	Aug 11, 2016	41523	

Unchanging working age

Another category of working age anomalies are instances where the working age is the same across two or more calendar dates. This does not immediately suggest an error, insofar as the engine simply may have been unused between these two dates. Table 4 provides a list of these anomalies.

Based on feedback from Kinross, the working ages were interpolated.

NEXT STEPS

At the current stage of the project, all responses from Kinross have been integrated into the data. Thus, the data are now ready for EXAKT analysis.

Identifying any critical variables (i.e. condition monitoring data) that appear to affect the engine hazard will enable building a proportional hazards model. Adding cost information will enable optimization of replacement decisions.

Table 4 List of instances with identical working ages on multiple dates

Unit #	Date 1	Date 2	Working age
801	July 7, 2014	July 21, 2014	102068
	Dec 22, 2014	Jan 5, 2015	104652
	May 19, 2015	Jun 1, 2015	107082
	Nov 6, 2015	Nov 25, 2015	110638
	Jun 11, 2016	July 5, 2016	114918
802	Dec 23, 2014	Jan 6, 2015	105177
	Oct 20, 2015	Nov 6, 2015	110699
	Nov 12, 2015	Nov 25, 2015	111203
	May 14, 2016	June 14 2016	114945
803	May 12, 2014	May 26, 2014	98332
	Dec 2, 2014	Dec 8 2014	101870
	Feb 23, 2015	Mar 2, 2015	103127
	Apr 6, 2015	Apr 15, 2015	103872
	Aug 3, 2015	Aug 10, 2015	106107
	Feb 22, 2016	Mar 21, 2016	110368
	Jun 6, 2016	Jun 30, 2016	111973
	Sep 13, 2017	Oct 9, 2017	121538
804	May 10, 2016	Jun 9, 2016	114991
805	Dec 25, 2014	Jan 1, 2015	105370
	Mar 26, 2015	Apr 10, 2015	106763
	May 28, 2015	Jun 12, 2015	107906
	Aug 26, 2015	Sep 4, 2015	109771
807	Aug 13, 2014	Sep 3, 2014	103073
	Jan 28, 2015	Feb 11, 2015	105858
	Apr 1, 2015	Apr 15, 2015	106819
	Apr 22, 2015	May 6, 2015	107115
	May 13, 2015	May 27, 2015	107534
809	Nov 24, 2014	Dec 8, 2014	94981
	Mar 30, 2015	Apr 13, 2015	96888
	May 25, 2015	Jun 1, 2015	97826
	Jun 21, 2017	Jul 19, 2017	113327
810	Apr 9, 2014	Apr 16, 2014	91494
	May 28, 2014	Jun 11, 2014	92379
	May 28, 2014	Jun 18, 2014	92379
	Apr 21, 2015	Apr 28, 2015	97812
	May 20, 2015	Jun 3, 2015	98385
	Sep 9, 2015	Sep 12, 2015	100766
	Nov 4, 2015	Nov 6 2015	101835
811	Apr 22, 2015	May 6, 2015	96584
	May 27, 2015	Jun 1, 2015	97329
	Oct 27, 2015	Nov 12, 2015	100593
	Nov 6, 2017	Nov 7, 2017	115514
812	Feb 4, 2015	Feb 11, 2015	96785
	Apr 30, 2015	May 6, 2015	98219
813	Jan 25, 2016	Feb 16, 2016	64442
814	Nov 13, 2013	Nov 14, 2013	49896
	Nov 24, 2013	Dec 5, 2013	50038
	May 1, 2014	May 8, 2014	53475
815	Aug 9, 2012	Aug 21, 2012	42577
	Feb 14, 2013	Feb 28, 2013	45840
	May 9, 2013	May 23, 2013	47111
	Oct 24, 2013	Nov 7, 2013	49889
	Nov 20, 2014	Nov 27, 2014	47097
	Jan 24, 2017	Feb 15, 2017	71903
816	May 14, 2015	May 29, 2015	34214
	Apr 2, 2016	Apr 25, 2016	40806
817	Nov 21, 2013	Dec 5, 2013	22695
	Apr 23, 2015	Apr 30, 2015	31992
	May 23, 2016	Jun 19, 2016	39982

TORONTO HYDRO INVESTMENT SPIKE SMOOTHING AND STEADY-STATE ANALYSIS

GARY (JIAYUE) WANG, PEY STUDENT

OVERVIEW

This report discusses the collaboration between Toronto Hydro and C-MORE on investment spike smoothing and steady-state analysis.

INVESTMENT SPIKE SMOOTHING AND STEADY-STATE ANALYSIS

Toronto Hydro has a large variety of physical assets in the distribution system; each type of asset has a large and dynamic population. The company has a replacement program(s) for each category of asset. All assets are subject to replacements, reactively and/or proactively, due to failures, failure risks, aging, upgrades, legal obligations, etc. Assets of the same type often have similar useful-life or optimal replacement time (explained in later section). Historically, large projects/initiatives have installed assets in a short period of time (or in a cyclical nature of a period of years), which can contribute to spikes in renewal expenditures in the future. Due to capital budget constraints, not all assets can be replaced at or before the optimal/expected replacement time, which further contributes to the spikes in renewal expenditures.

Toronto Hydro is interested in developing a model/strategy that allows planners to smooth out renewal expenditure spikes under time constraints and/or capital budget constraints while maintaining relatively high cost-efficiency.

Further, Toronto Hydro is interested in a steady-state analysis model/strategy that is able to examine the following:

- When system/program/asset expenditures will reach a steady-state, given:
    - Short-term forecasted expenditures before steady-state;
    - Estimated expenditure level in steady-state;
    - Estimated asset demographics in steady-state (i.e. the portion of the demographics that is past the useful life for the asset class).

DATA ANALYSIS

The research will first develop the model/strategy based on a few types of assets, then generalize the model/strategy and allow it to perform on a system-level/aggregated-level.

Optimal Replacement Time (ORT)

One of Toronto Hydro's investment models computes the optimal replacement time (ORT) using replacement cost and failure cost of assets. The ORT suggests the most cost-efficient time to replace an asset, given its asset value and consequences of failure.

Asset Type #1: Underground Transformers (UG TX)

Toronto Hydro has approximately 20,000 underground transformers in the distribution system (Source: Toronto Hydro Current-State Analysis). Based on the age demographics of underground transformers, the ORT approach suggests that about 25% of underground transformers are either at ORT or past ORT, and therefore should be replaced immediately for cost-efficiency purposes (call it "initial replacement"); see figure 1.

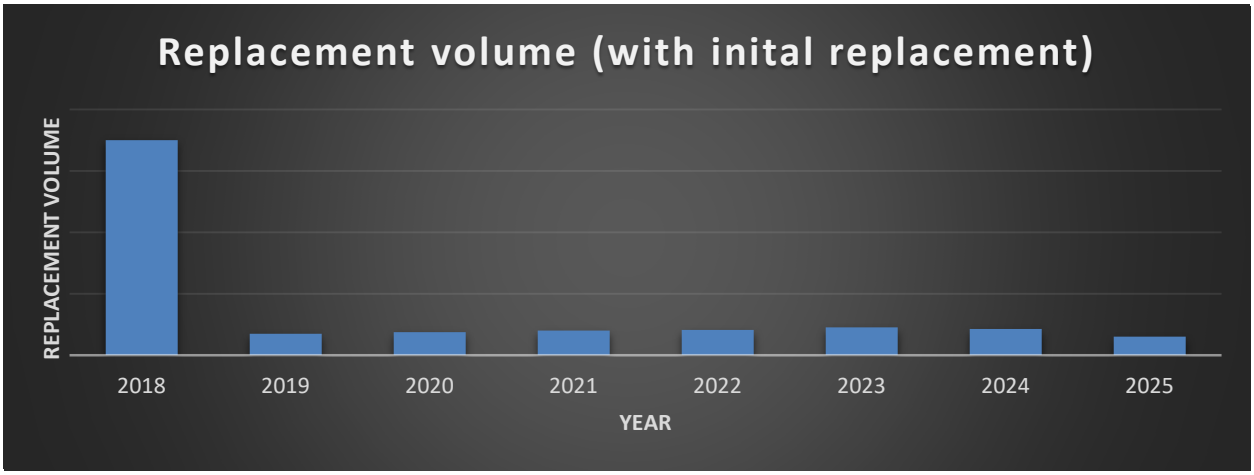


Figure 3 Underground Transformer Initial Replacement Amount Suggested by ORT (Source: Toronto Hydro Current-State Analysis)

After the suggested "initial replacement", the annual replacement volume ranges between 400-700 units per year. However, there are several spikes requiring replacement of up to 1,500 units in the next 100 years.

The initial replacement and replacement spikes are not feasible due to capital budget constraints. The objective of this project is to determine a smoothing plan to avoid replacement and investment spikes.

TECHNICAL REPORTS: PRINCESS MARGARET HOSPITAL

PREDICTING THE RELIABILITY OF LINEAR ACCELERATORS BY ANALYZING THE TRENDS AND CORRELATIONS OF FLATNESS OVER TIME

MOZAM S. SHAHIN AND DANIEL M. DUKLAS,
U of T UNDERGRADUATE STUDENTS

BACKGROUND

The following is a digest of the thesis submitted in conformity with the requirements for the degree of BASc, Department of Applied Science and Engineering, University of Toronto. The last progress report on the thesis is included in the C-MORE Report for December 2017.

ABSTRACT

The purpose of this thesis is to predict when linear accelerators (LINACs) need to be serviced through the analysis of flatness. Flatness measures the uniformity of the electron beam which ensures that patients receive the optimal amount of dosage by LINACs. It is vital that the patient receives the correct dosage. This is because if the patient receives too little or too much dosage it could be fatal. One of the measures of even dosage to a localized cancer tumour is field flatness. Our main goal was to predict flatness of linear accelerators.

To meet the goal, data from 11 linear accelerators in Princess Margaret Cancer Care Centre were analyzed. The linear accelerators were tested daily on the following two photon energies: 6 MV and 18 MV. This research focused on 6 MV and omitted the 18 MV because only the 6 MV data were complete. The team also obtained maintenance logs, as well as data about input and set parameters for the linear accelerators.

The data were then analyzed and it was determined that flatness was related to three different components:

1. The usage of the machine
2. Set parameter Gun_aim_I_Set_value
3. Maintenance and replacement of the parts

From there, two different types of models were created to predict the flatness of the machines:

1. A dynamic regression model that displays the short term trend when the R2 is higher than 0.25 and the P-value is less than or equal to 0.05.
2. Multiple static regression models with an R2of up to 0.7346 (depending on the machine).

These two models combined will help predict flatness more accurately. It is recommended that the dynamic model be used for short term predictions, whereas the static models be used for long term predictions and feature analysis.

INTRODUCTION

A linear accelerator is a device commonly used in cancer treatment. It is estimated that nearly fifty percent of all cancer patients receive some form of radiation therapy via a linear accelerator[1]. Linear accelerators (LINAC) emit electrons that, when aimed at the cancerous tissue, kill the cancer and potentially surrounding tissue.

LINAC radiation is emitted through an electron gun. Electrons are accelerated through the gun in the direction of the patient's cancer cells. This treatment is referred to as external beam radiation[2].

The correct dose (amount of energy) is determined by a radiation oncologist with the potential help of a radiation dosimetrist and a medical physicist. Once the correct dose is determined, the patient is treated by LINAC with the specific dose [3].

Over time the components of the LINAC machine start to deteriorate. This deterioration results in subpar dosage delivery. Deterioration can be caused by many factors. One of the factors is electron gun performance. To prevent subpar dosage delivery, LINACs undergo daily quality assurance tests.

One of the daily quality assurance tests is the flatness test. The flatness test measures the uniformity of the electron beam [4]. Uniform dosage is vital to killing all the cancer cells equally. Non-uniform dosage may either kill healthy tissue, or not kill cancerous tissue, or both.

Purpose

The purpose of this thesis is to predict when linear accelerators (LINACs) need to be serviced through the analysis of flatness and its correlation to different input parameters. If the serving time is not well predicted it can result in downtime or for the potential that the linear accelerator be used on a patient when it is supposed to (Note: there are existing preventative measures in place to prevent this). A well operating linear accelerator is important to ensure that patients receive the optimal dosage by the linear accelerator treating them.

Flatness of Linear Accelerators

Initially, to calibrate the LINAC's optimal flatness, the machine undergoes a water test. This test consists of using the electron gun to radiate electrons into a water tank. The water tank simulates the human body. The amount of radiation, along with its flatness, is measured.

Once the machines are calibrated, they are tested every morning by a sensor strip. Ideally, a water tank would be used to test the machines. However, it is not feasible to do so, because the water tank test usually takes a long time to complete. Instead, sensor strips measure the maximum and minimum intensity emitted by the electron beam. Figure 1 shows a diagram of a sensor strip measurement.

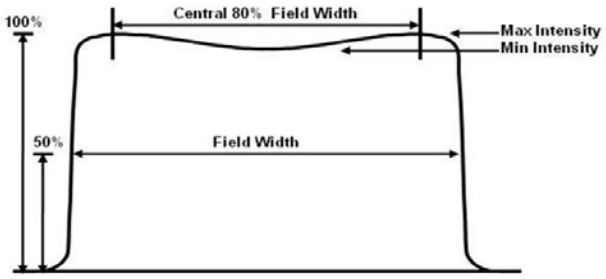


Figure 1: Flatness Measurement [5]

The field flatness is then calculated using the following equation:

$$Flatness = \frac{D_{\max} - D_{\min}}{D_{\max} + D_{\min}} \times 100\%$$

"D_{max} and D_{min} are the maximum and minimum dose along the profile within the core 80% of the field size."[6]

Each machine at Princess Margaret Cancer Centre has a slight different optimal flatness calibrated by the water tank because the machines have components that are installed or replaced at different times.

Further details on

- Characteristics of machines and testing protocol,
- Statistical analysis: Time series plots of the flatness,
- Maintenance Intervals for the Machines,
- Correlation between electron gun parameters and flatness,
- Correlation between other parameters and flatness, and
- Conjectures/Hypothesis on flatness trend and correlation with parameters,

can be found in the December 2017 C-MORE report. Here we give an overview of the final analysis and modeling. Complete thesis can be obtained upon request.

FEATURE SELECTION

Once exploratory data analysis was completed, variables were identified that would be useful in predicting flatness through more rigorous methods, using expert advice and correlation analysis.

From the analysis performed, it was determined that the all the parameters had some correlation to flatness but most had a very weak relationship. Whether the parameter is a significant predictor of flatness will be determined through building regression models and seeing how well the parameters predict flatness.

BUILDING A PREDICTION MODEL BASED ON CORRELATION ANALYSIS

With the significant parameters identified, the next step was to create a prediction model. The approach was to start with a simple model and then to build upon it. The first model was built using all the non correlated related parameters such as Gun I mean, Gun I standby Mean, Gun V standby Mean, Bal Set Val and Gun Aim I Set Value. Finally, an extra parameter labeled Flatness.ID was added to signify the date of the flatness. The first date that the flatness was taken was labeled 1 and the second day was labeled 2.

The parameters were inputted into Minitab and a stepwise regression was applied to them. The recommended regression model was in the form of:

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$$

The initial model chose only the Flatness.ID as a significant feature. Flatness.ID was added to signify the date since installation. For example, on the first day the machine was installed the Flatness.ID was 1. On the second day, the Flatness.ID was 2, and so on.

When we inputted the data in Minitab and conducted a stepwise regression on it, the final model had an R^2 of ~29%, which is a weak relationship. The Flatness.ID was significant with a p-value much less than 0.05. The final suggested equation was:

$$\text{Flatness} = 25.41 - 0.05935 \cdot \text{Flatness.ID} + 0.000041 \cdot \text{Flatness.ID}^2$$

Observations

From building the initial model, there were some interesting insights that were gained about the nature of the data. It became apparent that time was a significant factor in determining the flatness. In fact, time was the only parameter to be deemed significant.

IMPROVEMENTS

From the information of the original regression model and its shortcomings, the following changes were made to address the weakness of the model:

Extracting Additional Information from Data

Given the significance of the variable time on flatness that was identified in the initial observations, a variable that tracked the number of operating days since last maintenance was performed. This variable indicated how many days it had been since the last maintenance was performed. This variable does not count days on which the machine did not run (usually weekends), as the variable was created specifically to track machine usage.

Backfilling Data

There was a sparse amount of data for the parameter values because some of the parameters, changed manually by technicians, were only recorded when the values changed. When trying to perform a regression analysis, it became apparent that the amount of data being used was limited the number of days that the parameter was recorded for. In order to compensate for this, the parameter data were backfilled. Backfilling consisted of setting the values to the last saved value for the days that were missing parameter data.

Flatness	Gun Aim I Set value
4.4	811
4.75	
4.51	
4.59	
4.57	811

Flatness	Gun Aim I Set value
4.4	811
4.75	811
4.51	811
4.59	811
4.57	811

Figure 23: *Left*: Before backfilling. *Right*: After backfilling.

Multiple Models

The assumption made when creating the original model was that the behaviour of the machines were the same. That is, that the flatness followed the same trends. It was a reasonable assumption considering that the machines have the same model number and manufacturer, and usage was similar for most machines. However, after running the regression analysis on multiple machines, the machines seem to have different models that they performed well with. This suggested that using a few models instead of one may help better predict flatness.

DYNAMIC CURVE FITTING

In order to better predict the future flatness, a more dynamic approach was considered. This approach would entail fitting a regression model, given the last 20 days of data. (Any length of time can be substituted for 20 days.)

Method

Similar to the static approach, a linear regression model was chosen as the method to predict the flatness. The difference was that it was constantly being updated to fit the latest data. Also, it relied on a subset of the data, as will be explained.

Approach

From the static model, the gun aim I parameter was shown to be a statistically significant predictor of flatness. So this parameter was considered when designing the dynamic model. It can be observed that typically when the gun aim I parameter changed in value so did the flatness, which suggested some sort of maintenance was performed on the electron gun (see Figure 25). This assumption was used to determine what data to include.

Model

The model looked at the data of any particular machine since the last time maintenance had been performed. Then, it would fit a regression model on the data and make predictions for the next few days. This was not ideal, since regression models are not good at making future predictions when there is no general model, but it can help facilitate decision making about the trend of the data and approximate when the flatness will cross the warning threshold.

Output of the model

The following is the dynamic model prediction plot:

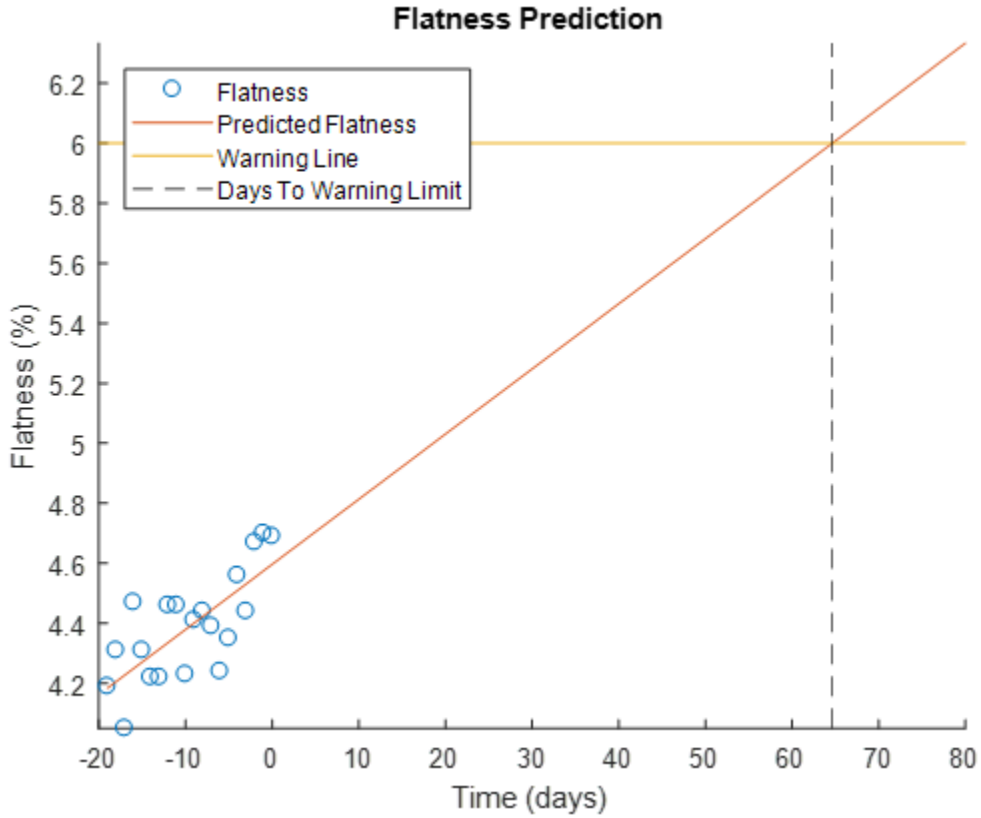


Figure 24: Plot of the flatness prediction (Machine: EA08)

In this specific example, the dynamic program predicts that the flatness will reach the warning limit of 6% in about 63 days. By using this information, the maintenance team will know that they will need to monitor the LINAC within that time frame.

Strengths & Weakness of Model

This model can more accurately predict when maintenance will be required in the future. However, the model prediction ability is a function of the amount of data available. Given that the data that can be used to create the dynamic model is the flatness since the last maintenance (i.e., the last time the gun aim I parameter changed) it can pose an issue in the days following a gun aim I parameter change. This is because a regression model built off only a few data points is unreliable.

UPDATED STATIC REGRESSION MODEL

The final regression model was created using the backfilling method outlined above for the Gun Aim I Set Value parameter. The following diagram shows the changes for the data:

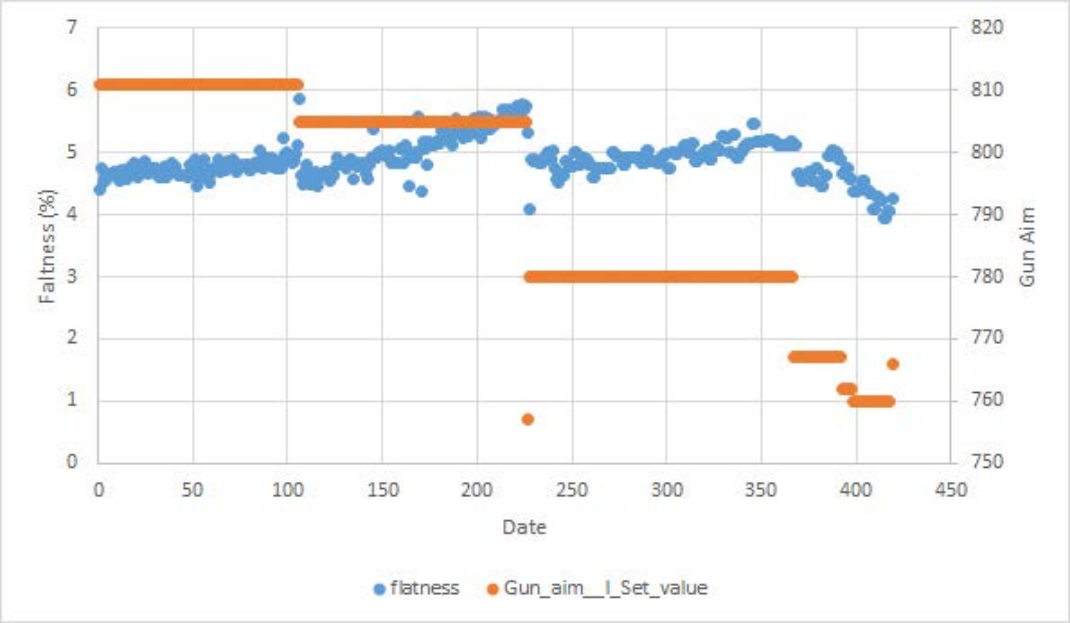
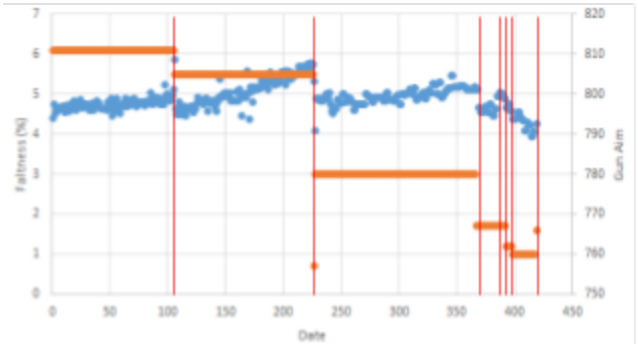


Figure 25: Plotting the Gun Aim parameter alongside a scatter plot of flatness

Once the data for the Gun Aim I Set Value parameter were successfully backfilled, two new features were created: (1), a parameter that determines the days since a maintenance interval, called “maintenance interval” and (2), a parameter that is binary and has the trend of the general flatness during the maintenance interval. “Up” would mean that the trend of the flatness is going up while down would mean that the overall trend during the maintenance interval is down. Figure 26 visually shows each parameter. On the left, the data show the parameters and features plotted against the flatness. On the right is an example of the values.



Flatness	Gun Aim I Set value	maintenance Interval	trend
4.4	811	1	up
4.75	811	2	up
4.51	811	3	up
4.59	811	4	up
4.57	811	5	up

Figure 26: *Left:* Parameters plotted against each other. *Right:* Example of values

The team then ran the data into a multiple regression in Minitab. The model's R^2 value is about 54.14%. A value of this magnitude is classified as moderate to substantial. This means that the regression model is generally well behaved. The results also show a p-value much smaller than 0.05 for all the parameters. Thus all values are significant.

Second, the team subsequently examined the residual values to verify that the assumptions of the regression model were not violated. The residuals were normal, the probability - probability plot was linear and the residuals versus fitted did not show a clear trend. Therefore, the assumptions of the regression were met.

General Model for all Machines

The regression model was repeated for various other machines and the following equation was empirically identified for all machines:

$$Flatness = \beta_0 + \beta_1(Gun\_aim\_I) + \beta_2(Maintenance\ Interval) + \beta_3(Gun\_aim\_I)^2 + \varepsilon$$

The betas are different for different machines. One reason for this is that some of the intra machine parameters are different for each machine.

RECOMMENDED USAGE OF THE DYNAMIC AND STATIC MODELS

It is determined that the dynamic model is accurate within the maintenance intervals while the general regression model (static model) is slightly more accurate for the general long-term trend and is useful to conduct feature analysis. This is because the dynamic model only looks at local flatness points and makes a prediction on the local behaviour of the linear accelerator. Also, the dynamic model would not work as accurately for long term predictions as it makes a linear up or down prediction and that trend is only exhibited in certain scenarios in the short term. For the long term trend and prediction analysis, it is recommended that the general regression model (Section 7.14) be used for the linear accelerators. It can help approximate its general trend and can also be used to analyze the impact of each feature. it should be noted that because the R2is not very high thus it can be used to make approximate predictions and not exact predictions. Using both of these models in tandem will result in the best approximation for flatness.

RECOMMENDATION FOR FUTURE IMPROVED MODELS

To improve the model further, data need to be collected more frequently so that the flatness and the parameter values align. One of the major obstacles we faced was determining the value of a specific parameter on a given day. We were able to backfill data for only one parameter and could not justify it for others as their values change constantly. Our recommendation would be to reconduct this study once there are complete data for all the parameters. To accomplish this, we recommend that the technician save all the parameters once the flatness is measured in the morning. Our second recommendation is that if there is a need to improve the accuracy of the predictions, then other methods such as random forests could be used. Those results are not readily comprehensible, however they will produce a more accurate prediction.

REFERENCES

[1]M. Baker, "Medical linear accelerator celebrates 50 years of treating cancer", *Stanford University*, 2017. [Online]. Available: http://news.stanford.edu/news/2007/april18/med-accelerator-041807.html. [Accessed: 11- Nov- 2017].

[2]Mayo Clinic, "External beam radiation for prostate cancer - Overview", *Mayo Clinic*, 2017. [Online]. Available: https://www.mayoclinic.org/tests-procedures/external-beam-radiation-for-prostate-cancer/home/ovc-20204694. [Accessed: 11- Nov- 2017].

[3]R. (ACR), "LINAC (Linear Accelerator)", *Radiologyinfo.org*, 2017. [Online]. Available: https://www.radiologyinfo.org/en/info.cfm?pg=linac. [Accessed: 11- Nov- 2017].

[4]F. Khan, *The physics of radiation therapy*, 3rd ed. Philadelphia, Pa.: Lippincott Williams & Wilkins, 2010, p. 211.

[5]C. Able, C. Hampton, A. Baydush and M. Munley, "Initial investigation using statistical process control for quality control of accelerator beam steering", *Radiation Oncology*, vol. 6, no. 1, p. 180, 2011.

[6]M. Hossain and J. Rhoades, "On beam quality and flatness of radiotherapy megavoltage photon beams", *Australasian Physical & Engineering Sciences in Medicine*, vol. 39, no. 1, pp. 135-145, 2015.

[7] Manfredi, T. (2018). *Machine parameters related to beam flatness.*. [email].

MIE 490 CAPSTONE DESIGN: FINAL DESIGN SPECIFICATION (FDS)

TONGLIN JIN, YUHENG LIN, XUEHAN WANG, AND YUZE LI,
MIE STUDENTS

BACKGROUND

The Capstone Design Project Report from October 20, 2017, was presented in the December 2017 report. Here we present the Executive Summary of the final version of the project.

Project name and date: PM3-Princess Margaret, Hospital LINAC Project, March 23, 2018.
Client Name: Daniel Letourneau, Chi-Guhn Lee.
Project Supervisor: Chi-Guhn Lee.
Communication Instructor: Kazemi Yasamin.

EXECUTIVE SUMMARY

Treatment for cancer plays a vital role in modern world as the increasing number of cancer patients. Radiation therapy is one of the most useful way for cancer treatment nowadays. To ensure the safety of the patients, the maintenance of the gun machine using for the radiation therapy becomes extremely important. However, the current procedure of the maintenance in PMH is considered highly maintenance-intensive and discrepant. Based on the identified problems in the current maintenance approach, the design solution will utilize the machine-recorded parameters to propose servicing intervention recommendations for all gun units by analyzing periodically assessed radiation flatness scorings.

Several objectives were set for the design. For instance, the design should be cost efficient, and the design should provide the maintenance estimations at least 7 days prior to servicing intervention, with accuracy. In addition, key constraints were also identified and met. First, the design must be compatible with the current platforms for the LINAC system as required by the client. Second, the design must indicate all the gun units which caused the change in operational performance. Moreover, there are numerous users and stakeholders associated with this project. For instance, the equipment maintenance team is the primary user.

In order to make the maintenance procedure more efficient, the team decided to develop a mobile app that can monitor the flatness parameters in all gun units. The detailed interface of the design is provided in Section 2.1. In order to support the front-end User Interface, the team developed the Data Prediction Model Application (DPMA) as the backend technology through defining key data-fitting algorithms to implementing the algorithms under Matlab environment. The machine's quality control data is periodically imported, merged and managed as a relational database management system. To accurately predict the operating status for all LINAC units, four common modern machine-learing approaches were used to train and predict the data. After continuous training, LASSO Linear Regression and Classification Regression Tree (CART) are implemented into DPMA as the best two fitting models. The Linear Regression algorithm

yielded an R-square prediction accuracy of 0.315, which is expected to improve after accumulation of future importing LINAC data. The CART algorithm yielded an R-square prediction accuracy of 0.8959, which indicates it is highly accurate for determining categorical outcomes (unit fail/not fail). For specific implementation and detailed results of the DPMA application, please refer to Section 2.2 - 2.5.

However, there are also some limitations associated with the design. The most notable is the computational complexity of the design. As continuous data streams are accumulated throughout the time for machine learning purpose, the resulting computation time increases exponentially by the increment of the size from the data fitting relation. Therefore, the ongoing dedicated computational power to the DPMA could be insufficient in the near future depending on the size of the updated LINAC data stream. Additional hardware investment should be considered upon observation of insufficient computational power. Alternatively, the design can be configured to abandon previous historic data to sacrifice some accuracy for computational efficiency.

For better improvement on the user experience and system automation, the next step for the team is to build a complete performance suite which includes a middle end to form the complete ecosystem.

TECHNICAL REPORTS: STUDENT RESEARCH

APPENDICES

APPENDIX I: WATER QUALITY PREDICTION AT MTENDELI REFUGEE CAMP IN TANZANIA WITH HIERARCHICAL CLUSTERING AND CUSTOM ENSEMBLE REGRESSION MODEL

JANGWON PARK, UNDERGRADUATE MIE SUDENT

BACKGROUND

The following is a shortened thesis submitted in conformity with the requirements for the degree of BASc, Department of Applied Science and Engineering, University of Toronto, under supervision of Professor Chi-Guhn Lee. It is included here as an example of an application of **Machine Learning** methodology to a real life problem. Note that many of the figures have been omitted but can be obtained upon request.

ABSTRACT

Well-known first- and second-order decay models lack robustness and sometimes demonstrate poor performance in accurately predicting future chlorine concentrations of water samples in a similar refugee camp setting to Mtendeli. A two-stage custom algorithm was developed to address these issues. In the first stage, hierarchical clustering analysis divides the entire data into three clusters. Each cluster contains only similar water samples that are likely to decay at very similar rates, thereby creating three distinct decay rate classes: low-, medium-, and high-decay. The optimal parameters to hierarchical clustering are determined by feature selection algorithms and a heuristic optimality score based on cophenetic correlation coefficient and cluster size variance. In the second stage, a custom ensemble regression model is developed by averaging predictions from random forest and gradient boosted trees. Predictions are made in each of the three clusters separately. The results demonstrate that the model is effective for predicting future chlorine concentrations of water samples in the "low-decay" cluster with an average R^2 of 0.88 over 100 random trials, suggesting some promise of machine learning in resolving the issues introduced by traditional statistical models. However, in the other two clusters, the model performance is rather unsatisfactory. Significant improvements can be made in both clustering analysis and machine learning predictions by incorporating data on new features such as tap stand conditions, temperature differential, total organic carbon, and hours of direct exposure to sunlight, and by including more data both in general as well as within each water sample.

1. INTRODUCTION

1.1 Background and Research Gap

Many developing countries in Africa, including Tanzania, lack the infrastructure for the kind of piped water supply systems often seen in developed countries and therefore rely heavily on centralized batch chlorination [1]. Under this setting, refugees at Mtendeli camp must fetch water manually, enough to last a day or two at a time, from various tap stands where treated water is distributed. However, it is overwhelmingly difficult to accurately predict when this water may

become unsafe to consume due to varying climate conditions every day and many opportunities of water contamination in an often-unsanitary refugee camp environment. Predicting an acceptable water quality – measured by the concentration of free residual chlorine (FRC) [mg/L] – is therefore crucial to the refugees' health and livelihood.

Two studies in the past have tackled the same problem using statistical models in a similar refugee camp setting in Africa [1, 2]. These models are well-known first- or second-order decay models which describe the decay rate of a water sample i.e. the rate at which the concentration of FRC decreases. However, two issues exist at large: 1) the statistical models are sometimes highly inaccurate, and 2) the models lack robustness as the author admits that the model parameters may change vastly even for a small improvement in R^2. The implication of an inaccurate, highly sensitive model is an unreliable and possibly incorrect recommendation of initial chlorine dose to water batches. As such, new approaches must be considered to fill this research gap.

1.2 Objectives

The objective of this project is to build a robust model which accurately predicts future chlorine (FRC) concentrations using machine learning techniques. The advantage of using machine learning for prediction is that it does not make assumptions on how chlorine will decay prior to prediction as statistical models do; this is particularly reliable in a refugee camp setting as it is an open, dynamic system where samples that are identical in all aspects may still undergo different rates of chlorine decay on different days purely due to weather conditions or unanticipated human factors.

2. LITERATURE REVIEW

Predicting a numerical, continuous target is called regression in the context of machine learning. Studies on predicting water quality using machine learning are quite scarce and therefore only a few works are discussed here. As a result, this section will also concentrate on building the conceptual foundation on some specific methods of machine learning regression which may prove effective for this project.

2.1 Clustering Analysis

Clustering is a method by which water samples that are inherently similar are grouped together into the same cluster prior to any prediction of chlorine concentration. It achieves this by minimizing the intra-cluster distances and/or maximizing inter-cluster distances accordingly. Countless number of works have used clustering analysis to label and structure their data set in numerous ways. For this project, the idea is that if clustering analysis can discover and classify only the very similar samples together, then they must follow the same model and decay at (more or less) the same rate of change. Predictions on these samples, then, could be much more accurate.

Two of the most common clustering methods are discussed: k-means and hierarchical clustering. The following table summarizes the pros and cons of each method in general.

***Table 5. Pros and cons of k-means and hierarchical clustering methods* [3]**

K-means Clustering		Hierarchical Clustering	
Pro	**Con**	**Pro**	**Con**
Generally more efficient than hierarchical clustering	The number of clusters¸ *k*, is often difficult to determine without expert opinion.	Far greater number of options to specify regarding the combination of method and metric when forming clusters within MATLAB	Requires significant tuning and experimentation
	Far fewer options to specify regarding the combination of method and metric	Can visualize merging of clusters with dendrograms	Difficult to identify the best cutoff point to limit the number of clusters

2.1.1 Measures of the Quality of Clusters

With using **either** k-means or hierarchical clustering, a major challenge is to evaluate the quality of the clusters – how good is the current clustering analysis? In this section, two formal measures are proposed to assist in answering that question. However, the project will also develop a heuristic measure of the quality of clusters based on how many data points each cluster has. For example, to ensure that the training set in each cluster is sufficiently large when building machine learning models, it is desirable to have evenly balanced clusters. Therefore, cluster size variance will be a good heuristic measure to consider in addition to the following.

2.1.1.1 Cophenetic Correlation Coefficient (CCC)

In hierarchical clustering, it is desirable for inter- and intra-cluster distances to be proportional. One way to verify this is by computing the cophenetic correlation coefficient which, if close to 1, suggests that the data is well suited for clustering [3]. Alternatively, cophenetic correlation coefficient can be thought as the correlation between the dissimilarity between any two clusters with their inter-cluster distance [4, 5]; the greater the degree of dissimilarity, the greater the inter-cluster distance. A clustering algorithm would ideally preserve this notion all throughout the algorithm.

CCC is particularly useful when comparing all possible combinations of methods and metrics in hierarchical clustering against each other; the combination with the highest value produces the best clustering analysis on the data. [4] does exactly this to figure out the best clustering method for its unique data set. It compares four different methods – Single, Average, Complete, and Ward – with six different metrics including Euclidean, Standard Euclidean, Minkowski, Mahalanobis, Manhattan, and Cosine. Among the 24 different combinations, the Ward-Cosine combination proved optimal. However, the paper suggests that the optimal combination varies from data set to data set, and each project must implement its own to discover the optimal combination for themselves. [5] performs the same analysis and again chooses its own optimal combination based on the highest CCC produced among 63 different combinations. This provides further confidence in the application of CCC to determine the optimal clustering method for this project.

2.1.1.2 Inconsistency Coefficient

A common stopping criterion in merging clusters is to specify the maximum number of clusters. A more algorithmic approach to this is to use the **inconsistency coefficient**. For cluster A and B about to be merged, the coefficient is calculated as follows [3]:

$$i(A,B) = \frac{d(A,B) - \mu}{\sigma}$$

The distance metric *d(A, B)* is substituted by an appropriate metric i.e. Euclidean, Manhattan, etc. The parameters μ and σ are the mean and standard deviation of the distances in the previous merges. The equation effectively standardizes the merge about to be formed to the historical inter-cluster distances. Therefore, the greater the value of *i*(*A*, *B*), the worse, or the more "inconsistent", the merge is. A good clustering analysis often involves identifying the threshold on the inconsistency coefficient so that any potential merges whose inconsistency coefficients exceed this threshold are not created.

[4] again does exactly this analysis to determine the "right" number of clusters. Its approach is to iterate through multiple computations of the inconsistency coefficient on the same data set, each time with a different "depth" parameter. The depth parameter specifies how much historical inter-cluster distances one would consider when computing the inconsistency coefficient. The greater the depth, the higher the inconsistency coefficient will be for the merge in question since earlier merges were always formed between much closer clusters. The heuristic approach in [4] equates the depth parameter as the number of clusters to use, and graphically observes the trend or pattern in inconsistency coefficients with subsequent merges at each depth parameter. The general rule is to find the depth at which the graph appears in a more "compact" form and the maximum inconsistency coefficient does not rise as much, or slows down significantly. Many works relate to this particular paper to leverage its heuristic approach in their applications, and likewise, this project may perform a similar analysis.

2.2 Regression Trees

Unlike clustering analysis which is an unsupervised machine learning method, regression trees are a supervised, completely data-driven method of predicting the values of the target variable using decision trees. A decision tree uses a number of rules to divide the feature space into rectangles in a way that identifies regions having the most homogeneous responses to the features, capturing complex nonlinearities between them and the target variable [6]. There are largely two different algorithms which make use of decision trees for prediction: random forest and gradient boosted trees.

2.2.1 Random Forest

Random forest is an ensemble (collection) of many decision trees. Random forest typically makes use of a bagging algorithm. In a bagging algorithm, a subset of observations is randomly sampled with replacement from the full training set with equal weights [7]. By default, random forests develop deep trees, which are grown independently and in parallel [7]. Each tree's prediction ultimately casts a "vote" to decide on the final prediction based on the majority rule. Because each tree only uses a subset of the training set, random forests avoid overfitting the training data.

[8] uses random forest to predict the degree of nitrate pollution in groundwater in Southern Spain. Nitrates typically originate from inorganic fertilizers in farming and can pollute water when discovered in excess amounts. Though this is in a different context, the application is similar to this project in that both aim to evaluate the quality of water by using random forest and a bagging algorithm for regression. The authors conclude that random forest is a very promising predictive method for water research and performs better than linear regression across many criteria including mean squared error. However, the discussion is rather inconclusive since the only benchmark is linear regression and because it will be difficult to generalize this result to a non-agricultural context. This indicates a need to carry this research forward to a different context and compare random forest to a broader variety of other predictive models.

2.2.2 Gradient Boosted Trees

Gradient boosted trees are similar to random forests with one notable difference: it makes use of a boosting algorithm which assigns different weights to the observations when it randomly samples a subset of the training set. Rather than growing trees independently and in parallel, gradient boosted trees are grown sequentially. At each sequence, a new tree focuses on observations that the model did not predict well in the previous sequence by adjusting the weights [6]. The goodness of predictions is evaluated by a loss function with the most typical one being sum of squared errors (SSE). The goal of gradient boosted trees algorithm is thus to keep adding trees, whose predictions are ultimately averaged for the same observation, to minimize the loss function iteratively. Gradient boosting consists of weak learners, which are "shallow" trees in this case.

A specific study for using gradient boosted trees for water quality prediction does not exist. It is reasonably assumed that as they are promising since random forest, which is based on decision trees as well, was found to be promising for water quality predictions. With no prominent research in leveraging this technique for water quality prediction yet, this project may provide meaningful ground for the validity of its performance.

2.3 Support Vector Regression

Support vector regression (SVR) is a regression technique based on statistical learning theory and a structural risk minimization principle, which had great successes in nonlinear modeling [9]. It attempts to discover a predictive function based on data by minimizing the generalized error bound rather than the observed error as statistical models do.

[9] applies SVR to predict water quality in river crab habitats over time as measured partly by dissolved oxygen and partly by temperature. It achieves great success in accurately predicting and describing the changes in water quality with time in comparison to back-propagation (BP) neural networks. River crab habitats are open, nonlinear, dynamic and complex settings which makes this study particularly relevant to the context of this project. Additionally, [10] applies least squares support vector regression in likewise nonlinear, dynamic, and complex Liuxi River in Guangzhou, China and predicts water quality very accurately in comparison ARIMA models or BP neural networks, and further highlights the strength of SVR in generalizing an accurate function to describe often uncontrolled target values. Generalization of the findings of these two

studies is not practical as the research was conducted in a very specific environment. Therefore, application of SVR in this project may contribute to the current state of the field.

2.4 K-Nearest Neighbor

K-nearest neighbor (KNN) is known to be one of the simplest algorithms in terms of implementation which also proved very successful in many applications over the years. Traditionally, it is a non-parametric method (i.e. makes no assumptions on the function or the underlying distribution of data) for pattern classification based on the estimates of *k* nearest neighbors about the observation in question [11]. For regression, it works very similarly by averaging the estimates of the *k* nearest neighbors.

Research on using KNN for regression is relatively scarce compared to classification. Though it has been applied in various contexts from basal area diameter prediction to lithium-ion battery capacity prediction, regression applications for water quality prediction is extremely rare. [12] shares a limited insight into the performance of KNN for water quality classification. First, the setting is non-open, controlled water resources systems. Secondly, the quality prediction is a multi-class classification problem, which may be vastly different from regression in nature. Thirdly, thought KNN was found to work to some extent, its performance as measured by error rate was poorer compared to support vector machines or neural networks [12]. However, in consideration of the limitations and potential improvements KNN may achieve in a completely different setting such as the Mtendeli refugee camp, it may be sought as a candidate algorithm for making predictions.

3. DESCRIPTION OF THE DATA

The data used in the project are provided by Syed Imran Ali, an affiliate researcher at Doctors Without Borders. The entire data set contains 145 observations with a total of 22 features including initial chlorine concentration, initial temperature, type of container, etc. Each observation contains two data points – initial chlorine concentration recorded at time 0 and a final chlorine concentration recorded anywhere between 15 to 25 hours post-distribution. The data are expected to propose two major challenges for accurate prediction due to its small size:

- The regression model may not be robust because smaller data sets usually produce inconsistent results with each trial as the impact of data partitioning is much greater.
- Having only two data points may render some machine learning techniques ineffective such as SVR since the generalized model could likely be a straight line connecting the two points. This would be far from accurately describing chlorine decay, which usually exhibits a first- or second-order exponential decay.

4. METHODOLOGY

The custom algorithm developed for predicting the second (also the final) chlorine concentration in each water sample largely consists of two sequential stages: hierarchical clustering and custom ensemble regression model using average predictions from random forest (RF) and gradient boosted trees (GBT). In the first stage, hierarchical clustering requires three parameters – method, metric, and a subset of features. The first two parameters shall be determined using a

heuristic optimality score computed based on cophenetic correlation score and cluster size variance while the subset of features is determined by three different feature selection algorithms: neighborhood component analysis (NCA), lasso regularization, and correlation criteria. Detailed discussion on this is presented in the next section. In the second stage, R^2 and regression error characteristic curves (RECC) are used to evaluate and select the best predictive model.

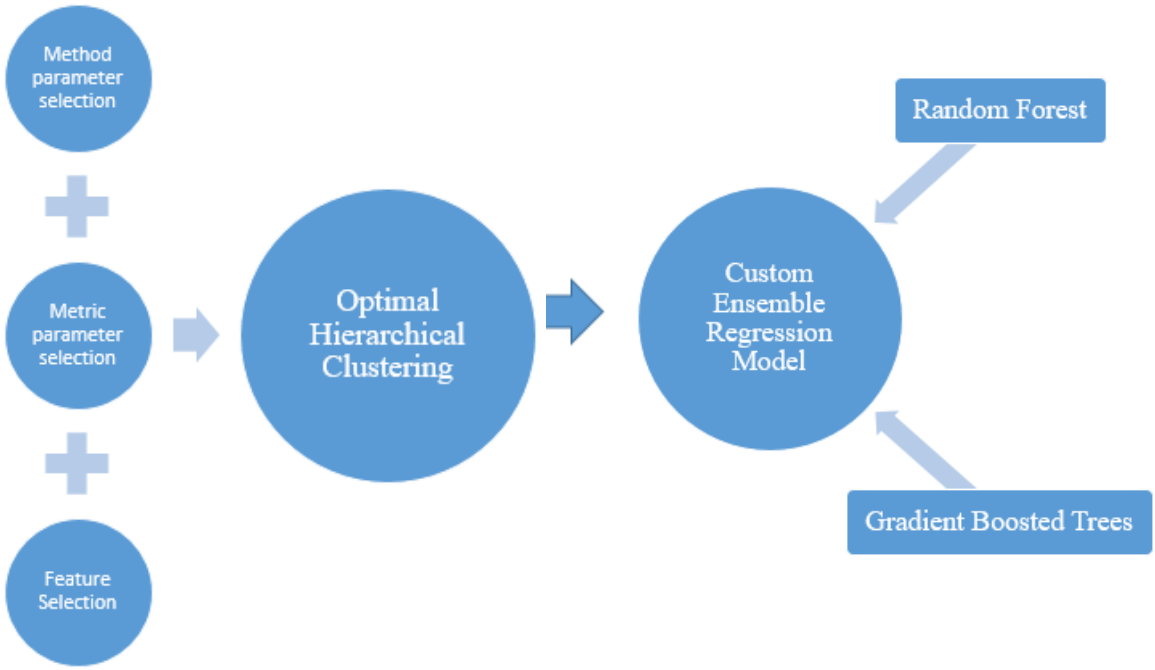


Figure 4. Intuitive diagram of the custom algorithm developed for the project

4.1 Feature Selection

The objective of feature selection is often three-fold: improving prediction accuracy of predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data [11]. The purpose of feature selection in the context of this project lies most heavily in the third objective where feature selection may assist hierarchical clustering in maximizing its ability to identify consistent patterns only among water samples with similar decay behavior; here, the "underlying process" shall refer to the unique decay rate that each observation undergoes. Though expert domain knowledge may suffice if present, more algorithmic approaches can help select significant features, or more appropriately, exclude undesirable features that are highly correlated with others and/or add noise or bias to predictions. For robustness, three different algorithms are considered: neighborhood component analysis (NCA), lasso regularization, and correlation criteria.

NCA is an embedded method which minimizes prediction error of an underlying predictive algorithm with an additional regularization term [12]. Functionally, it is equal to k-nearest neighbors (KNN) and makes use of stochastic nearest neighbors, and the regularization term

represents the sum of squared weights of each feature and therefore is conceptually equal to ridge regression [12]. Therefore, NCA can extract features that contribute to accurate predictions of chlorine concentration. In this project, only the feature weights that are greater than 0.01 are selected for the hierarchical clustering stage.

Similarly, lasso regularization is an embedded method which minimizes the sum of squared errors between observed data and predictions made by linear regression [13]. In lasso regression, the beta coefficients of the features are encouraged to be reduced to zero and thus is an effective method for feature selection [14].

Finally, correlation criteria describe Pearson correlation coefficient between each of the feature to the target variable i.e. second (final) chlorine concentration [11]. In this project, features whose absolute correlation coefficient are less than 0.15 are not considered as inputs to the hierarchical clustering algorithm.

4.2 Hierarchical Clustering

Hierarchical clustering is a promising and reliable approach to group "similar" observations. Similarity is a concept of distance – the closer the two observations are, the more similar they are. A proper distance function is uniquely defined by every method-metric combination supported in MATLAB.

Table 6. List of methods and metrics supported by MATLAB in hierarchical clustering

Method	Metric
Single	Euclidean
Complete	Standard Euclidean
Average	Squared Euclidean*
Centroid*	Mahalanois
Ward	Cityblock (Manhattan)*
Median	Chebychev
Weighted	Minkowski
	Hamming*
	Cosine*
	Correlation*
	Jaccard
	Spearman

*Method and metrics supported by MATLAB for k-means clustering.

When a group of water samples all exhibit a similar decay behavior, thus demonstrating a consistent pattern among them, accurately predicting chlorine concentrations may become significantly easier. Therefore, the **goal of the clustering analysis is to group the entire data into a few clusters such that each cluster may only have the samples will all likely undergo the same rate of decay**. Feature selection algorithms proposed in the previous section only exists to complement this processing by advising which set of features may be provided as inputs to the clustering algorithm. If successful, it can be shown that hierarchical clustering is a simple, but powerful data-structuring pre-analysis to complement predictive algorithms.

The procedure of hierarchical clustering analysis in this project can be summarized in the following steps:

1. Input features selected from the feature selection algorithms and a specific method-metric parameter into the hierarchical clustering algorithm. Specify the maximum number of clusters to three.
2. Compute the score for the provided method-metric parameter by averaging the cophenetic correlation coefficient (CCC) of the algorithm and normalized cluster size variance. Store this score.
3. Repeat steps 1 and 2 for all possible method-metric combinations.
4. Choose optimal method-metric combination corresponding to the highest score from step 4.

Relying solely on CCC may sometimes produce high unbalanced clusters where only one or two observations are found in a cluster while the rest are concentrated into another. Therefore, an arithmetic average is calculated between CCC and cluster size variance to acquire good quality clusters while also ensuring roughly evenly sized clusters.

Finally, while k-means clustering can be much quicker to implement, hierarchical clustering presents more combinations of methods and metrics to experiment with in MATLAB to discover the optimal approach for the specific data in question. This outweighs the advantage of k-means clustering as the accuracy of the algorithm is ultimately prioritized over its efficiency.

4.3 Custom Ensemble Regression Model

4.3.1 Data Preprocessing

Prior to predicting chlorine concentrations with the custom ensemble regression model, some data preprocessing is beneficial to improve the performance as much as possible. Two topics are discussed here – data partitioning/validation and feature engineering approaches.

4.3.1.1 Data Partitioning/Validation

70% of the entire data set is generally used for training and the rest for testing. However, small datasets are particularly sensitive to how the data is partitioned between the training and testing sets and consequently may yield significantly varying estimations of model accuracy [13]. To make the model more robust and help generalize it for any data set, 5-fold cross validation is implemented. In a *K*-fold cross validation, the training set is partitioned in to *K* equal portions, exactly one of which will always serve as a validation set when training the model; this is repeated *K* times each with a different validation set [13]. Cross validation is therefore an effective approach which allows one to make the best use of scarce data.

4.3.1.2 Feature Engineering

Feature engineering approaches can be supplementary tools to boost model performance. Two of the most common methods are feature scaling, which normalizes all numerical features to place their relative importance/influence on predictions on a common scale, and principal component analysis (PCA) to address undesirable correlation and noise among features which may harm model performance [15].

PCA is essentially a change of basis technique which computes the most meaningful basis to re-express a potentially noisy data [15]. It achieves this by projecting the original feature space to

lower dimensions which are effectively a set of orthogonal directions, or principal components, which explain the most variability of the data. In this project, by selecting the first several principal components which cover 99% of the variability of the data, the essence of the data is retained while a completely new set of "features" are derived.

4.3.2 Candidate Algorithms

There are several candidate algorithms. In this project, the following are considered: k-nearest neighbors (KNN), support vector regression (SVR), random forest (RF), gradient boosted trees (GBT), and a custom ensemble model which averages predictions of any combinations of the techniques. The custom ensemble model eventually achieves the highest performance and is therefore pursued. The details of how each algorithm works is discussed in the literature review section. In this section, only the strengths and limitations are briefly discussed.

Table 7. Outline of each predictive algorithm's strengths and limitations

	KNN	SVR	RF	GBT
Strengths	Non-parametric	Structural risk minimization More robust than statistical models	Non-parametric Handles outliers and overfitting well	Non-parametric
Limitations	Influenced heavily by binary variables Sensitive to outliers	Relies on having many observations for accuracy	Generally slower run-time performance Easy to lose interpretability with large numbers of trees	May not handle overfitting well Easy to lose interpretability with large numbers of trees

4.3.3 Selection Criteria

The following two common quantitative measures exist for selecting the "best" predictive model which be used to evaluate each model in the previous section: R^2 and regression error characteristic curve. R^2 is intuitively the measure of performance of a model relative to a simple average line for predicting chlorine concentrations. By default, the R^2 of this benchmark is 0 and any model whose accuracy is positive indicates that it performs better than the benchmark; however, the goal is to get as close to 1 as possible. Since R^2 is a very well-known measure of performance in regression, only the regression error characteristic curve will be discussed.

4.3.4 Regression error characteristic curve (RECC)

For classification problems, receiver operating characteristic (ROC) is a widely used, powerful visualization and comparing the predictive power of classification models. Regression error characteristic curves (RECC) generalize ROC to regression and therefore serve the same purpose for comparing regression models, which has gained popularity in many studies over the years [18]. On the *x*-axis of a RECC, error tolerance is plotted often computed by squared errors while

on the *y*-axis the accuracy of the regression model is plotted from 0 to 1. Naturally, as the error tolerance increases, the accurate of the model nears and ultimately reaches 1. Qualitatively, the RECC of a good model tends towards the top left corner of the graph; quantitatively, this means it achieves a high area under the curve (AUC). A great AUC is highly correlated with a high R^2 value and therefore both will be used to evaluate model goodness.

5 RESULTS AND DISCUSSIONS (figures excluded)

5.1 Feature Selection

The following summarizes the significant features as identified by each of the three feature selection algorithms:

Features	NCA	Lasso Regularization	Correlation Criteria
Initial FRC concentration*	✓	✓	✓
Initial TRC concentration*	✓	✓	✓
Initial temperature*	✓	✓	✓
Jerrycan	✓	✓	✓
Bucket	✓		✓
Other container			
White		✓	✓
Green			
Blue	✓		
Yellow	✓		
Orange			
Pink			
Red			
Container opacity	✓		✓
Container covering	✓		
Container cleanness			
Same container	✓		
Same water	✓		
Container fullness*	✓	✓	✓
Method of drawing	✓	✓	✓
Container outside			
Time elapsed*			

*numerical variables (the rest are binary)

There are fundamental limitations with each of these algorithms. For instance, the underlying predictive model in NCA is KNN while for lasso regularization, it is linear regression. As these may not be effective models for making accurate predictions in this project, the results here are only lightly considered as suggestions. With some trial and errors, the following features are selected as inputs to the clustering algorithm, which are also highlighted in the table above: initial FRC concentration, initial temperature, jerrycan, bucket, container covering, and container cleanness.

5.2 Hierarchical Clustering

With these input features, and each of the 84 method-metric combination from Table 3, the hierarchical clustering is ready to be evaluated. As mentioned, CCC and cluster size variance are averaged to output a final score for each combination.

Figure 5. CCC scores of each method-metric combination.

Figure 6. Cluster size variance for each method-metric combination

CCC scores do not generally differ as significantly as cluster size variance does across all method-metric combinations. Therefore, cluster size variance scores have a lot of impact on determining the final, optimal method-metric parameters for the data. Prior to averaging both scores, the cluster size variances are first normalized by dividing by the largest observed variance and then inverted to allow the smallest variance to have the highest score (lower the variance, the better). At the end of this analysis, it is discovered that the **Ward** method with the **Euclidean** metric are optimal.

Table 8. Result of Ward-Euclidean hierarchical clustering analysis

Evaluation Criteria	Result of Clustering		
	Cluster 1	Cluster 2	Cluster 3
Number of samples	49	43	53
Range of time elapsed [h]	8.4	8.5	8.8
Mean initial chlorine [mg/L]	0.7	0.8	0.7
Mean decay rate [mg/L/h]	0.33	0.46	0.61

While the first three evaluation criteria are not distinct enough to distinguish a cluster from another, the mean decay rates of the clusters are sufficiently different. Hypothesis testing on equal means and medians using the two-sample t-test and Wilcoxon rank sum test reveals that the clusters are indeed statistically different based on the decay rates at the 5% significance level. This is an important finding in this project as it is a working, supporting proof of the objective of the clustering analysis to some extent. In conclusion, this supports the idea that **the data set can possibly be divided based on decay rates into three clusters based on three distinct decay rate categories – namely low-, medium-, and high-decay clusters**.

However, the clustering analysis needs to be improved significantly to allow for more precise clustering to produce more distinct clusters i.e. their mean decay rates should be further apart than now. Table 4 presents the best result that cannot be improved as of now. This implies that the current set of features are not informative enough to cluster the data more precisely. However, new features that contain suggesting information about a water sample's decay rate can potentially improve clustering analysis if available. Some insightful conjectures are proposed below to support this claim.

Conjecture 1: Tap stands distribute water samples with significantly varying decay rates, but the same tap stand tends to distribute samples with consistently similar decay rates.
The current data set has 52 unique tap stands based on their unique ID's. Each tap stand contributes anywhere from one to five samples/observations to the data set. No features on tap stand conditions such as age or cleanness exist. As a result, it is currently assumed that 1) all tap stands are equally capable of distributing water samples of all kinds of decay rates from low- to high-decay, and 2) tap stands do not significantly vary from each other in affecting decay rates. However, an analysis into the mean decay rates of the samples from each tap stand hints otherwise.

Figure 7. Mean decay rates of each tap stand in decreasing order

The height of each bar represents how many samples are recorded from that tap stand in the current data set, and the black trendline is the mean decay rates of all samples from that tap stand. When arranged in the decreasing order of this mean decay rate, the figure shows most of the tap stand only tends to distribute water samples of one kind of decay rate. For example, all three observations from tap stand 2 are labelled "high rate" while four of five observations from tap stand 48 are labelled "low rate". Therefore, it can be expected that the next sample collected from these two tap stands are very likely to exhibit high decay and low-decay respectively. Furthermore, contrary to the current assumption, tap stands do vary significantly from each other in what kind of decay rates they each tend to produce.

Figure 8. Distribution of mean decay rates of the tap stands (left) and normality check of the distribution using (right)

In fact, the distribution of the mean decay rates of the tap stands tend to follow a very close normal distribution as presented in figure 4 with the mean of 0.48 mg/L/h and standard deviation 0.21 mg/L/h. Coefficient of variation can be used as an effective descriptive statistic to indicate the dispersion of the data, which in this case is 44%. In general, this is a very large dispersion and thus helps dispel the assumption that tap stands do not vary significantly from one another in affecting decay rates. In conclusion, this conjecture reveals that tap stands may distribute observations with largely varying decay rates, and that a tap stand may consistently distribute samples of one decay rate category.

Conjecture 2: Temperature differential correlates positively with decay rates
Temperature is widely known to accelerate decay rate when higher. However, the temperature distribution in the current data set is not diverse and is in fact discrete.

Figure 9. Distribution of initial temperatures of all observations

Therefore, the effect of temperature on decay rate cannot be studied effectively. In addition, the positive effect of higher temperatures on decay rates is reasonable in closed systems where the temperature is maintained. In an open, dynamic setting such as the Mtendeli refugee camp, where the temperature can fluctuate throughout the day, the initial temperature alone may not be sufficient even if the distribution of the temperatures is continuous. An alternative measure is temperature differential – the difference between the temperatures at various points in the future

and the initial temperature. Fortunately, the current data set has recorded a final temperature for each observation. These, however, are not used for clustering or predictions since they are not known in advance.

Figure 10. Box plots of temperature differentials of each cluster for the 3-cluster approach (left) and for the 5-cluster approach (right). The clusters are arranged in increasing order of mean decay rate.

The green diamond indicates the mean temperature differential of each cluster. With both the 3-cluster and 5-cluster approach, it can be shown that the mean temperature differential correlates positively with the mean decay rate of each cluster. This is reasonable since an observation will have decayed at a faster pace if it experiences a higher change in temperature. Such an analysis with the initial temperature alone does not reveal the same correlation.

In addition to conjectures 1 and 2, other features which are closely related to decay rates are total organic carbon (TOC) and hours of direct exposure to sunlight. TOC is an estimate of the cleanness of water itself which reacts with chlorine. Therefore, it is expected to accelerate decay rate when found in greater amounts of concentration. As it may require a specific equipment to measure TOC, some papers have suggested a linear relationship between TOC and chlorine demand, thereby possibly allowing one to compute TOC without measuring it [19]. As for hours of direct exposure to sunlight, it is intuitively to understand that it will accelerate decay rate. The current data set has a feature "container outside" which attempts to capture the same information. However, it overlooks the fact that each day experiences varying degrees of cloudiness and rain which renders this feature ineffective.

Though the hierarchical clustering results lack precision now, further analysis found promising results and potentials for their improvement. To verify the idea that clustering based on the decay rate is indeed complementary to machine learning prediction, the rest of this paper will assume that the algorithm is improved and therefore have clustered directly on decay rates for further illustrations.

5.3 Custom Ensemble Regression Model

Directly clustering on the decay rate leads to three precisely divided data sets:

Table 9. Comparison of results between "perfect" clustering and empirical clustering

Evaluation Criteria	"Perfect" Clustering			Result of Clustering		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Number of samples	49	43	53	49	43	53
Range of time elapsed [h]	8.0	9.1	9.9	8.4	8.5	8.8
Mean initial chlorine [mg/L]	0.7	0.7	0.7	0.7	0.8	0.7
Mean decay rate [mg/L/h]	-0.14	-0.40	-0.84	-0.33	-0.46	-0.61

By the same method presented in the previous section, it was discovered that the **Weighted** method with the **Squared Euclidean** metric performed the best among all other options. The

decay rate used in the "perfect" clustering scenario was **linearly approximated**. The only reason for the linear approximation is because each observation in fact has only two data points; attempting to fit first- or second-order decay models despite this fact led to a highly concentration distribution of decay rates as shown:

Figure 11. Distribution of decay rates as approximated by first-order (left) and second-order (right) models

The fact that the vast majority of the observations have more or less the same decay rate makes it impractical to cluster the data into three groups based on three distinct decay rate categories. Again, this is due to lack of data points in each observation, and therefore linear approximation was pursued.

Selection of the "best" regression model for predicting chlorine concentrations in each cluster is decided by each candidate model's performance as measured by R^2 and RECC. For convenience, only the results for the "low-decay" cluster are presented.

Figure 12. Regression error characteristic curves of different models for the low-decay cluster

Table 10. Average R^2 over 100 trials for each model for the low-decay cluster

	KNN	SVR	GBT	RF	Ensemble
R^2	-0.2	-0.08	0.82	0.87	**0.88**

The benchmark is simply a horizontal, average line to predict chlorine concentrations and therefore has an R^2 of 0. The ensemble model, which averages the predictions of GBT and RF, not only achieves the highest AUC in the RECC, but it also has the highest R^2 over 100 trials. For these reasons, the ensemble model is selected as the optimal model. Nine trials are shown below for demonstration purposes:

Figure 13. Predicted (red) vs. actual (blue) chlorine concentration for the low-decay cluster

Data preprocessing approaches such as 5-fold cross validation, feature scaling, principal component analysis led to a moderate improvement in model performance after implementation.

The same degree of accuracy, however, are not seen in the other two clusters. While the average R^2 for the medium-decay cluster was 0.71, it was 0.08 for the high decay cluster over 100 trials, which is only slightly better than the benchmark.

Table 11. Average R^2 over 100 trials using the custom ensemble regression model

	Low-Decay Cluster	**Medium-Decay Cluster**	**High-Decay Cluster**
R^2	0.88	0.71	0.08

Figure 14. Predicted (red) vs. actual (blue) chlorine concentration for the high-decay cluster

This is likely due to the fundamental limitation imposed by the linear approximation of the decay rate. In reality, chlorine in water does not decay linearly; it undergoes a form of exponential decay which cannot possibly be observed accurately with a crude linear approximation. This can be avoided by collecting more data points in each observation. Once there are at least four or five data points, the decay rate can be estimated much more accurately for that observation. Clustering analysis and machine learning prediction with these observations are then expected to improve dramatically.

It is also concluded that the reason predictions are more accurate in clusters with lower mean decay rates is primarily because the observations in the low-decay cluster, for instance, have much smaller difference between their initial and final chlorine concentrations, which often make predictions easier.

6 CONCLUSION

A custom algorithm was implemented to pick the optimal parameters for hierarchical clustering that would best group the entire data into three clusters based on three distinct decay rate categories, namely low-, medium-, and high-decay rates. The parameter selection process consisted of three feature selection algorithms – NCA, lasso regularization, and correlation criteria – as well as score calculations using CCC and cluster size variance. It is expected that significant improvements to the precision of clustering can be achieved by incorporating data on new features such as tap stand conditions (age, cleanness, …), temperature differential, total organic carbon, and hours of direct exposure to sunlight.

A custom ensemble regression model which averages predictions from RF and GBT was selected to be the optimal model based on its R^2 over 100 trials and its performance on the RECC plot. Clustering analysis proved effective for the low-decay cluster, achieving an average R^2 of 0.88, but not for the other two clusters. It is concluded that this is most probably due to linearly – and quite crudely – approximating the decay rates. With more data points in each observation, decay rates can be estimated more accurately using well-known first- or second-order decay models. Not only will this improve machine learning predictions, it is also expected to improve the precision of the clustering analysis.

Other improvements include collecting more observations in general. The current data set is too small in the context of machine learning (145 observations). Larger training sets are bound to make the model more robust and potentially improve model accuracy as well.

7 REFERENCES

[1] S. I. Ali, S. S. Ali and J.-F. Fesselet, "Effectiveness of emergency water treatment practices in refugee camps in South Sudan," *Bulletin of the World Health Organization*, 2015.

[2] S. I. Ali, "Study Report: Evidence Based FRC Targets for Centralized Chlorination in Emergencies," Medecins Sans Frontieres, 2017.

[3] T. Chan, *Clustering,* Toronto: University of Toronto, Department of Mechanical and Industrial Engienering, MIE465 Analytics in Action, 2018.

[4] B. Uragun and R. Rajan, "The discrimination of interaural level different sensitivity functions: Development of a taxonomic data template," *BMC Neuroscience,* vol. 14, pp. 114-134, 2013.

[5] S. Saracli, N. Dogan and I. Dogan, "Comparison of hierarchical cluster analysis methods by cophenetic correlation," *Journal of Inequalities and Applications,* vol. 1, pp. 203-211, 2013.

[6] J. Elith, J. Leathwick and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology,* vol. 77, pp. 802-813, 2008.

[7] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[8] V. Rodriguez-Galiano, M. P. Mendes, M. J. Garcia-Soldado, M. Chica-Olmo and R. Luis, "Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulunerability: A case study in an agricultural setting (Southern Spain)," *Science of the Total Environment,* vol. 476, pp. 189-206, 2014.

[9] S. Liu, H. Tai, Q. Ding, D. Li, L. Xu and Y. Wei, "A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction," *Mathematical and Computer Modelling,* vol. 58, pp. 458-465, 2013.

[10] X. Yunrong and J. Liangzhong, "Water Quality Prediction Using LS-SVM with Particle Swarm Optimization," in *Knowledge Discovery and Data Mining,* Moscow, 2009.

[11] L. E. Peterson, "K-nearest neighbor," Scholarpedia, 21 February 2009. [Online]. Available: http://scholarpedia.org/article/K-nearest_neighbor. [Accessed 8 April 2018].

[12] F. Modaresi and S. Araghinejad, "A Comparative Assessment of Support Vector Machines, Probabilistic Neural Networks, and K-Nearest Neighbor Algorithms for Water Quality Classification," *Water Resources Management,* vol. 28, no. 12, pp. 4095-4111, 2014.

[13] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[14] W. Yang, K. Wang and W. Zuo, "Neighborhood Component Feature Selection for High-Dimensional Data," *Journal of Computers,* vol. 7, no. 1, pp. 161-169, 2012.

[15] T. Chan, *Linear Regression,* Toronto: University of Toronto, Department of Mechanical and Industrial Engineering, MIE465 Analytics in Action, 2018.

[16] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* vol. 58, no. 1, pp. 267-288, 1996.

[17] J. Shlens, "A Tutorial on Principal Component Analysis," Cornell University, 2014.

[18] J. Bi and K. P. Bennett, "Regression Error Characteristic Curves," in *Proceedings of the Twentieth International Conference on Machine Learning,* Washington DC, 2003.

[19] L. Yee, M. Abdullah, S. Ata and B. Ishak, "Dissolved organic matter and its impact on the chlorine demand of treated water," *Malaysian Journal of Analytical Sciences,* vol. 10, no. 2, pp. 243-250, 2006.

8 APPENDICES (graphs excluded)

8.1 Appendix A: Exploratory data analysis I – basic statistics of select features

Some basic statistics of select features are presented here.

Table 12. Basic statistics of time elapsed and linearly approximated decay rates

	Max	Mean	Min	Range
Time Elapsed [h]	25.9	20.3	15.8	10.1
Decay Rate [mg/L/h]	1	0.47	0	1
Initial chlorine [mg/L]	1.27	0.71	0.32	0.95
Final chlorine [mg/L]	1.12	0.38	0	1.12

8.2 Appendix B: Exploratory data analysis II – correlation analysis

Exploratory data analysis on the non-clustered (entire) data set reveals close to no correlations among the features, which partly motivated the pursuit of clustering analysis to begin with.

One would expect, for instance, to observe high decay rate for a water sample which started out with a high initial chlorine concentration if all water samples truly followed the same model. However, this is not found to be true. In conclusion, no obvious or intuitive correlations are discovered among the features, which motivated clustering analysis.

APPENDIX II: REINFORCEMENT LEARNING WITH MULTIPLE EXPERTS: A BAYESIAN MODEL COMBINATION APPROACH

MICHAEL GIMELFARB, PHD PRE-CANDIDATE

BACKGROUND

This work is the joint effort of Mr. Michael Gimelfarb, Dr. Chi-Guhn Lee and Dr. Scott Sanner in the department of MIE, University of Toronto. The work is complete and has been submitted to the Conference on Neural Information Processing Systems (NIPS) 2018.

DEFINITIONS

In reinforcement learning, we are interested in finding an optimal plan or policy to take over a future time horizon, when there is inherent randomness in the underlying state variables. Such stochastic processes are best modeled as *Markov decision processes (MDP):*

- S is the state space
- A is the action space
- $P = \{P_a : a \in A\}$ is a collection of transition probabilities, where P_a is the matrix of whose element at s, s' is denoted $p(s'|s, a)$
- γ is a discount factor in [0,1]
- $R: S \times S \times A \to \mathbb{R}$ is the reward function, in which $r(s, a, s')$ describes reward obtained when in state s, action a is chosen, and then a transition occurs to state s'.

We define a *policy* μ as a sequence of functions $\mu_0, \mu_1 \ldots$ from states to actions. Given an arbitrary MDP (S, A, P, γ, R) and policy μ, we can compute the discounted total infinite-horizon reward as

$$V^{\mu}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s\right]$$

where s_t is the state at time t evolving according to M and a_t is the action taken in state s_t, e.g. $\mu_t(s_t)$. More importantly, we are interested in finding a policy with the largest expected reward over all states, $V^*(s) = \sup_{\mu} V^{\mu}(s)$. In this case, $V^{\mu^*} = V^*$ where μ^* is called the *optimal policy*.

There are many algorithms for solving MDPs directly, including value and policy iteration (see, e.g. [Be95]). However, these algorithms all suffer from the curse of dimensionality and are no longer practical for large-scale MDPs. In order to address the curse of dimensionality, we instead maintain a table of values for each state-action pair $Q(s, a)$, which is defined as

$$Q(s, a) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_0 = a\right].$$

Typically, the Q-values are computed in an online (iterative) framework

$$Q(s,a) = Q(s,a) + \alpha[R_t - Q(s,a)] \qquad (1)$$

where α is a learning rate parameter and the second term $R_t - Q(s,a)$ is an error term which is the difference between our estimate of the future reward R_t and the reward estimate from the Q-table. There are many ways in which we can estimate R_t. The well-known *SARSA* (state-action-reward-state-action) update uses the one-step bootstrap returns

$$R_t = R_t^{(1)} = r(s_t, a_t, s_{t+1}) + \gamma Q(s_{t+1}, a_{t+1}),$$

where s_{t+1} is sampled according to P_{a_t} and a_{t+1} is chosen according to some exploration policy. One such policy is *epsilon-greedy* in which we select $a \in \text{argmax}_a\, Q(s,a)$ with high probability, or a random action with low probability. A more sophisticated estimation procedure, called *TD-lambda*, defines the rewards recursively as

$$R_t^{\lambda} = r(s_t, a_t, s_{t+1}) + \gamma\left[(1-\lambda)Q(s_{t+1}, a_{t+1}) + \lambda R_{t+1}^{\lambda}\right] \qquad (2)$$

where λ is a positive tuning parameter (see, e.g. [SB98]) After a fixed number of episodes, we can obtain the best policy by choosing the entry in $\{Q(s,a), a \in A\}$ with the highest value.

THEORY OF REWARD SHAPING

While (1) and (2) are often useful in practice, training can take a long time if the rewards are relatively sparse. In other words, if $r(s,a,s') = 0$ for a large number of elements, then the errors $[R_t - Q(s,a)]$ will often be relatively small and the Q-values will be updated relatively infrequently. In order to help speed up the learning process, it is often useful to "shape" the original reward function R into another function R' defined by

$$r'(s,a,s') = r(s,a,s') + F(s,a,s')$$

by introducing a shaping function F. If F has many non-zero elements, then so too will R' and learning can be accelerated.

However, it is necessary to exercise caution in defining the shaping function, because it is possible to alter the reward structure in such a way that the optimal policies for the original MDP are no longer optimal for the new MDP. Fortunately, it has been shown that the only class of shaping functions which preserves policy invariance is the *potential-based shaping function*.

Theorem 1 [NDS99]. *Let $M = (S, A, P, \gamma, R)$ be an MDP and let $M' = (S, A, P, \gamma, R + F)$ be the MDP after reward shaping. Then any policy which is optimal for M is also optimal for M' (and vice-versa) if and only if F is potential based, that is,*

$$F(s,a,s') = \gamma\Phi(s') - \Phi(s)$$

for some function $\Phi: S \to \mathbb{R}$.

Furthermore, and crucially in our analysis, the policy invariance property has been extended for *dynamic reward shaping* where

$$F(s, t, s', t') = \gamma\Phi(s', t') - \Phi(s, t)$$

and the potential is allowed to depend on time [DK12].

BAYESIAN REWARD SHAPING

The decision maker is provided with advice from N experts in the form of potential functions $\Phi_1, \Phi_2, \ldots \Phi_N$. Such advice could come from heuristics or guesses, from similar solved problems and can be analytic or computational. Unfortunately, in practice, this advice could often be contradictory or contain numerical errors, in which case it could hurt convergence. In order to make optimal use of the expert advice during the learning process, the agent should be able to learn which expert to trust as more information becomes available, and act optimally on this knowledge by applying the techniques in the previous section.

The two main approaches to incorporating multiple models in a Bayesian framework are *Bayesian model averaging (BMA)* and *Bayesian model combination (BMC)*. Roughly speaking, taking experts as hypotheses, BMA converges asymptotically toward the optimal *hypothesis*, while BMC converges toward the optimal *ensemble*. The model combination approach has two clear advantages over model averaging: (1) when two or more potential functions are optimal, it will converge to a linear combination of them, and (2) it provides an estimator with reduced variance. In this section, we show how BMC can be used to incorporate imperfect advice from multiple experts into reinforcement learning problems, all with the same space and time complexity as TD-learning.

In the general setting of Bayesian model combination, we interpret Q-values for each state-action pair $Q(s, a)$ as random variables, and maintain a set of past return observations D and a multivariate posterior probability distribution $P(q|D)$ over Q-values. We also maintain a posterior probability distribution $\pi(w)$ over the (N-1)-dimensional probability simplex S. Here, weight vectors w are interpreted as categorical distributions over experts; such a mechanism will allow us to learn the optimal distribution over experts, rather than a single expert. In the following subsections, we show how to maintain each of these distributions over time, but here we show how to use them for the general RL problem.

Given state s and action a at time t, we are able to show that the return under Bayesian model combination $R(s, a)$ is

$$R(s, a) = \sum_{i=1}^{N} E[w_i] E\left[q_{s,a} \middle| i\right]$$

where w_i are the components of the weight vector w. This result is both intuitively and computationally pleasing, and shows that ***the total return can be written as a linear combination of individual return ``contributions" from each expert model, weighted by the expected posterior belief that the expert is correct***. We now show how each of these two expectations can be computed.

ASSUMED DENSITY FILTERING

Starting with prior distribution π_t at time t over the simplex S and given new data point *d*, we would like to perform a posterior update using Bayes' theorem. Unfortunately, we show in our work that the exact posterior update is analytically intractable, so we need to use an approximation.

Assumed density filtering (ADF) or *moment matching* projects the true posterior distribution onto an exponential subfamily of proposal distributions by minimizing the KL-divergence (a notion of separation) between the posterior and the proposal distribution. We note that an excellent exponential family proposal distribution for our problem is the multivariate Dirichlet distribution with vector parameter α and density function

$$f(w) = \frac{Gamma(\sum \alpha_i)}{\prod Gamma(\alpha_i)} \prod \prod w_i^{\alpha_i - 1}, w \in S.$$

We apply moment matching by solving a system of non-linear equations $\frac{\alpha_i}{\alpha_0} = m_i$ and $\frac{\alpha_1(\alpha_1+1)}{\alpha_0(\alpha_0+1)} = s_1$ where α_i are the new Dirichlet parameters, $\alpha_0 = \alpha_1 + \cdots + \alpha_N$ and m_i are s_1 the means and variance of w_i and w_1, respectively, which we can obtain in closed form through Bayes' rule in terms of the old parameter $\alpha_{i,t}$

$$m_i = \frac{\alpha_{i,t}(e_i + e * \alpha_t)}{(e * \alpha_t)(\alpha_{0,t} + 1)}$$
$$s_1 = \frac{\alpha_{1,t}(\alpha_{1,t} + 1)(2e_1 + e * \alpha_t)}{(e * \alpha_t)(\alpha_{0,t} + 1)(\alpha_{0,t} + 1)}$$

where *e* is the evidence or the vector of probability of observing the data *d* given the optimal expert is *i*. It is not difficult to obtain a closed form and simple solution

$$\alpha_i = m_i \left(\frac{m_1 - s_1}{s_1 - m_1^2} \right).$$

This leads to an efficient O(N) algorithm for posterior updates in Algorithm 1.

Algorithm 1 PosteriorUpdate($\boldsymbol{\alpha}_t, e$)

for $i = 1, 2 \ldots N-1$ **do** ▷ Compute posterior moments
  $m_i \leftarrow \frac{\alpha_{t,i}(e_i + \boldsymbol{e}\cdot\boldsymbol{\alpha}_t)}{(\boldsymbol{e}\cdot\boldsymbol{\alpha}_t)(\alpha_{t,0}+1)}$
$s_1 \leftarrow \frac{\alpha_{t,1}(\alpha_{t,1}+1)(2e_1 + \boldsymbol{e}\cdot\boldsymbol{\alpha}_t)}{(\boldsymbol{e}\cdot\boldsymbol{\alpha}_t)(\alpha_{t,0}+1)(\alpha_{t,0}+2)}$
$\alpha_{t+1,0} \leftarrow \frac{m_1 - s_1}{s_1 - m_1^2}$ ▷ Compute $\boldsymbol{\alpha}_{t+1}$
for $i = 1, 2 \ldots N-1$ **do**
  $\alpha_{t+1,i} \leftarrow m_i \alpha_{t+1,0}$
$\alpha_{t+1,N} \leftarrow \alpha_{t+1,0} - \sum_{i=1}^{N-1} \alpha_{t+1,i}$
return α_{t+1}

It remains to show how to compute the evidence e.

Following the *Bayesian Q-learning* framework, we model Q-values for each state-action pair as independent Gaussian distributed random variables. Since the best potential function should be most representative of the optimal value function, we model Q-values $ given the best expert *i* as Gaussian random variables

$$q_{s,a}|i \sim N(\Phi_i(s), \sigma^2)$$

where σ^2 is tuned online from return data during training. Using these observations, our expected return at time t

$$R(s,a) = \sum_{i=1}^{N} E[w_i] E\left[q_{s,a}|i\right] = \frac{\sum_{i=1}^{N} \alpha_{i,t} \Phi_i(s)}{\sum_{i=1}^{N} \alpha_{i,t}} = \Phi(s),$$

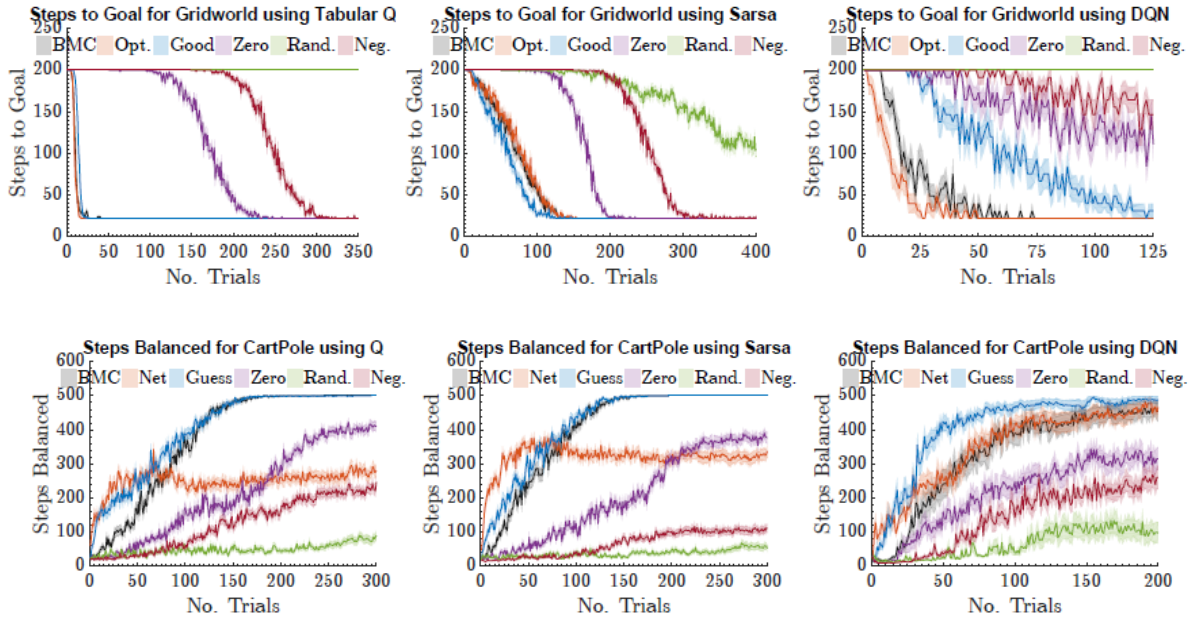
which represents the combined expert that is used for reward shaping. This leads to the complete Bayesian reinforcement learning algorithm with reward shaping in Algorithm 2. Here, TrainRL is a general procedure for training the agent using a generic reinforcement algorithm using the shaping rewards F determined from $R(s,a)$.

Algorithm 2 RL with Bayesian Reward Shaping

initialize $\boldsymbol{\alpha} \in \mathbb{R}_+^N$
for *episode* $= 0, 1 \ldots M$ **do** ▷ Main loop
  $\hat{\Phi} \leftarrow \frac{\sum_{i=1}^{N} \Phi_i \alpha_i}{\sum_{i=1}^{N} \alpha_i}$ ▷ Pool experts and compute shaped reward
  $F(s,a,s') \leftarrow \gamma \hat{\Phi}(s') - \hat{\Phi}(s)$
  $(R_t, s_t)_{t=1\ldots T} \leftarrow \texttt{TrainRL}(F)$ ▷ Perform one episode of training
  for all (R_t, s_t) **do** ▷ Posterior update
    update $\hat{\sigma}^2$ and compute $\boldsymbol{e}$
    $\boldsymbol{\alpha} \leftarrow \texttt{PosteriorUpdate}(\boldsymbol{\alpha}, \boldsymbol{e})$

EMPIRICAL RESULTS

We tested our algorithm (Algorithm 2) on two problem domains, the classical CartPole domain and a grid-world with 5 sub-goals, using tabular Q-learning, tabular Sarsa, and the Deep Q-Learning algorithm [MK15] first applied by DeepMind to solve Atari games. This is implemented using Python with Keras and Tensorflow backend. The number of steps required to reach the goals for grid-world and number of steps the pole is balanced are plotted below (all hyper-parameter configurations and architectures are given in our working paper, as well as the particular selection of experts, available on request)



What we find overall is our method is able to converge to the right combination of experts, even when the experts are noisy or imprecise, regardless of the function approximation or deep architecture. This approach can help improve robustness of reinforcement learning algorithms, and in particular in deep Q learning applications, and speed up convergence even when it is not clear which expert to trust.

REFERENCES

[Be95] Bertsekas, Dimitri P., et al. *Dynamic programming and optimal control*. Vol. 1. No. 2. Belmont, MA: Athena scientific, 1995.

[DK12] Devlin, Sam, and Daniel Kudenko. "Dynamic potential-based reward shaping." *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[MK15] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.

[NDS99] Ng, Andrew Y., Daishi Harada, and Stuart Russell. "Policy invariance under reward transformations: Theory and application to reward shaping." ICML. Vol. 99. 1999.

[SB98] Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. Vol. 1. No. 1. Cambridge: MIT press, 1998.