IICWG-XIX Breakout Summary Report
Big Data, Artificial Intelligence, and the Role of the Ice Analyst

Introduction

An important emerging issue for ice services is that of so-called "big data". There was a good introductory session on this topic at the 2017 meeting in Hobart where we concluded that "big data" is here and has major impact in our community, now and in the years to come. It will affect not only in the way we handle and analyze data, but also on the ice information products and services we provide to mariners.

It is a topic that calls for deeper analysis, understanding, and discussion, including investigations on where it makes sense for us to collaborate.

At IICWG-XIX in September 2018, the participants broke into 8 small groups to discuss how Ice Services are addressing the issues of big data and machine learning by considering the following questions:

Q1. How will ice analysis and information provision change in the future?
- What role will ice analysts performing in 5-10 years?
- What can analysts do to improve the confidence in ice information products? Is this something ice services are responsible for?
- How does/could machine learning contribute to modern ice charts?
- How could AI and machine learning tools be implemented in an ice service?
- Will it be possible to provide ice information multiple times a day?

Q2. Considering the explosion in the volume of data becoming available, what is the best place for data analysis?
- How will analysts develop trust in black-box solutions?
- Must data analysis be done in the ice service to be synergistic with other information (e.g. user reports)?
- What are the benefits and challenges of using cloud based systems for data storage and handling?
- How can we streamline the data toolbox?
- What is the future for Sigrid 3 and standardization for operational ice products?

Q3. How will ice surveillance improve over time?
- Can ice services integrate information with drones, reconnaissance and other sources?
- The OODA loop is something that is always in place for operators. How do we fit in here from an operations perspective?

Q4. Should ice charts work towards being more targeted to specific activities?
- Whose responsibility is it to provide this type of data for this type of incident?

Q5. Are there, or should there be, boundaries for commercial and government ice information services?

- Are they specific to different countries?
- Are there ways in which government and commercial ice services can work cooperatively for mutual benefit and to best serve end users?
- Is it better for everyone to be independent?

Discussion

Role of Ice Analysts/Forecasters vis à vis Machine Learning and Automation

All of the groups agreed that analysis and forecasting processes will become automated to greater and greater degrees. The following points were raised:

- There will be a transition from manual analysis to more automation. Efficiency will increase allowing organizations to produce more output faster within the same resources.
- There will always be a role for specialist analysts. However, their roles will change to focus more on dynamic, important, or difficult areas such as choke points, channels, fairways and port approaches. They will produce information for tactical decision making and deliver value-added information to automated products.
- Ice analysts/forecasters may become advisors with more interaction with users. Some participants felt that the role of the ice analyst should be to answer the phone and give tailored support.
- Analysts will also perform Quality Assurance tasks to ensure data and product quality. They will be supported by ML (and eventually Deep Learning) systems. It was noted that QA and certification of automated products may become issues that need to be addressed explicitly.
- As their roles evolve, analysts and forecasters will need new skillsets to maintain service excellence.
- Automated products produced by Machine Learning (ML) systems will help the analyst but there are constraints. Some participants felt that ML can only be as good as human analysts - it could save time but will not create better products. The availability of data to train ML systems is likely the largest stumbling block.
- It was pointed out that automated systems have not yet been shown to be capable. The timing of when that may become possible is very uncertain. Some think it is time to start preparing for a major change now while others believe we have 15-20 years before that is necessary.
- It is important that the science and technology teams building ML systems engage actively with ice analysts. Experience is essential. Better ground truth will be essential for validation.
- Ice forecasts are important. As we get better at forecasting, we will focus more on optimization for models and on communicating uncertainties in forecasts.

Cloud Computing as a Means of Handling the Explosion in the Volume of Data

There were two main camps in the cloud computing discussions. Proponents of the cloud raised the following points:

- Analysis should be moved into the cloud infrastructure with common virtual architectures that cross country borders. The cloud should be a repository for data.
- There could be a common data/algorithm architecture housed in a data center with products from every ice service in the same format.
- Black box processes could be performed by a third party – ice services don't have the time or resources for continuous development and maintenance.
- Standardized format and content is necessary to make borderless cloud computing work.
- With a common virtual center, training databases could be easily shared internationally.

However, cloud skeptics noted that:

- Cloud computing and storage are not free. Costs are not transparent and we don't what they will be like in future.
- There may be national security issues with cloud computing, particularly in government agencies. There could also be business continuity issues if the owner/operator of the cloud is not secure.
- The location for data analysis is situation-dependent and contingent on national standards and regulations.
- Developers of black box solutions must have an intimate understanding of the problems. Third party solutions will not be viable if developed in isolation.

One group predicted a scenario where, for the first 5 years, analysts will train the black box. In the next 5 years, they will watch the black box do the work. After that, they will forget how the black box functions!

It was noted that for users to develop trust in automatically produced products:

- R&D teams must show validation/verification of automated products to convince operators of their value, and
- Confidence maps would help develop trust in the systems and show human analysts where to pay additional attention to ensure product quality.

One group warned of the risk of falling into the "trough of disillusionment" if products are pushed to users before they are sufficiently robust, citing the IRIS routing system in the Baltic Sea that failed because users did not trust it.

The question about the future of SIGRID-3 was considered irrelevant and not addressed by most groups.

The Future of Ice Surveillance

Everyone agreed that ice surveillance will improve given the increasing number of platforms and channels of data giving more and more detailed observations. Drones will be used increasingly to confirm ice conditions in low confidence regions. However, more and more data presents challenges:

- Integrating information from multiple sources will require ML and black box processing.
- ML could support crowd-sourcing of ice observations.

- A global database needed. Ice services could put all of their observations into the cloud for universal access. We need to share data internationally.
- Standardized processes for reporting data to a black box in the cloud from all sources will be required.
- Filtering datasets and removing noise in black box solutions must be done with care to avoid the loss of information.

It was again noted that, while automated systems will be possible, we need to build user confidence in the products.

Should Ice Charts Be More Targeted to Specific Activities?

The groups discussed the idea of having ice charts that are more detailed and more tailored to support specific activities. The raised a number of points:

- Increased automation will allow ice analysts and forecasters to provide more focussed mission support, such as subset charts that focus on small areas, harbors, fairways, and channels.
- More tailored or more sub-regional area analyses and forecasts will depend on what users want and are willing to pay for.
- We could provide more parameters and more products but it is a challenge to avoid information overload. Users generally want simple, reliable, timely, and relevant products.
- Ships need a heads-up that they are moving into hazardous waters. The onus is on risk management.
- For general maritime safety, standard products should be free but enhanced products may be a paid service.
- Products must take communications bandwidth into consideration. There should be the capability to support users with low cost, low bandwidth communication systems as well as high-budget users with high bandwidth communications.
- Need to handle low bandwidth communications and let the user select what they want.
- Ice chart polygons were invented decades ago. Products of the future may look very different, such as information routinely delivered directly into users' decision support systems.
- There will always be a need for continuity in analyses – tailored products done in a discontinuous manner will not be adequate.
- Forecast products must look different from analysis products to avoid user confusion.

It was noted that, whatever shape future products take, the development process must include an interactive dialogue between users and ice analysts/forecasters, as well as between analysts, forecasters and machines.

Should There Be Boundaries for Commercial and Government Ice Information Services?

This question was asked in the on-line survey conducted before the meeting. It was ranked Low or Very Low priority by all responding groups. One breakout group spent the entire hour available and so they deemed it must be important. The following points were made:

- Government agencies have a responsibility to provide a basic service for maritime safety. That could extend to regional and sub-regional areas of interest such as choke points or port approaches.
- Government agencies should be authoritative source for ice information, look after Climatologies, and support Polar Code requirements.
- Government agencies should be the main source of data and use it to support commercial ice providers.
- Ice products that are more specific, such as tailored ship routes should be provided commercially but be regulated by the marine industry.
- Regardless of any boundaries that are established, infrastructure and regulation will be needed by both users and ice services.
- This question of boundaries and what government agencies can do commercially depends on country-specific legal frameworks.
- There can be tension between government and commercial organizations but there are good examples of successful public-private partnerships. Links between the two should be explored.
- Government should provide quality metrics – especially for automated products.
- While the best scenario from the user perspective is for all providers to share their information, this can be difficult when competing commercial entities need to protect their intellectual property.

One group proposed that there are lessons to be learned from the aviation industry. Airlines have their own weather forecasting teams to provide products tailored to their operations while still relying on government agencies for authoritative data and products. They outfit aircraft with automated sensors to great increasing the availability of in-situ data, which is in turn provided to government agencies for use in Numerical Weather Prediction, improving forecast quality or all. Could the marine industry ever evolve to this state?

And finally, it was suggested that the IICWG could bring the national ice services together into the International Ice Service.

Summary

In the wrap-up, everyone agreed that we all want to provide simple, reliable products for the mariners. Having a central repository for data along with a complete historical collection of ice charts would help towards that goal – but there are major obstacles to achieving that.

Ice services should share data among themselves as much as with clients. All services should have the same view of the world so they would not produce differing charts that confuse users.

And lastly, it would be useful for ice services to talk about who is going the way of automation in near future so experiences, developments, and results can be shared.

Appendix - Big Data Dictionary, version 0.1

Modified from https://campus.sagepub.com/blog/glossary-of-big-data-terms

Aggregation – a process of searching, gathering and presenting data

Algorithms – a mathematical formula that can perform certain analyses on data

Analytics – the discovery of insights in data

Anonymization – making data anonymous; removing all data points that could lead to identify a person

API - an application programming interface is a set of subroutine definitions, protocols, and tools for building application software

Application – computer software that enables a computer to perform a certain task

Artificial Intelligence – developing intelligence machines and software that are capable of perceiving the environment and take corresponding action when required and even learn from those actions.

Behavioural Analytics – analytics that informs about the how, why and what instead of just the who and when. It looks at humanized patterns in the da

Big Data - Big data is an all-encompassing term for any collection of data sets so large or complex that it becomes difficult to process them using traditional data-processing applications.

Brontobyte - A brontobyte is a measure of memory or data storage that is equal to 10 to the 27th power of bytes. There are approximately 1,024 yottabytes in a brontobyte. Approximately 1,024 brontobytes make up a geopbyte.

Classification analysis - a systematic process for obtaining important and relevant information about data, also meta data called; data about data.

Cloud computing – a distributed computing system over a network used for storing data off-premises

Clustering analysis – the process of identifying objects that are similar to each other and cluster them in order to understand the differences as well as the similarities within the data.

Comparative analysis – it ensures a step-by-step procedure of comparisons and calculations to detect patterns within very large data sets.

Complex structured data – data that are composed of two or more complex, complicated, and interrelated parts that cannot be easily interpreted by structured query languages and tools.

Computer generated data – data generated by computers such as log files

Concurrency – performing and executing multiple tasks and processes at the same time

Data aggregation tools - the process of transforming scattered data from numerous sources into a single new one.

Data analyst – someone analysing, modelling, cleaning or processing data

Database – a digital collection of data stored via a certain technique

Database Management System– collecting, storing and providing access of data

Data centre – a physical location that houses the servers for storing data

Data cleansing – the process of reviewing and revising data in order to delete duplicates, correct errors and provide consistency

Data custodian– someone who is responsible for the technical environment necessary for data storage

Data feed – a stream of data such as a Twitter feed or RSS

Data mining – the process of finding certain patterns or information from data sets

Data modelling – the analysis of data objects using data modelling techniques to create insights from the data

Data set – a collection of data

Data virtualization – a data integration process in order to gain more insights. Usually it involves databases, applications, file systems, websites, big data techniques, etc.)

Deep Learning: a subfield of machine learning, deep learning imitates the human brain by building artificial neural networks.

Exploratory analysis – finding patterns within data without standard procedures or methods. It is a means of discovering the data and to find the data sets main characteristics.

Exabytes – approximately 1000 petabytes or 1 billion gigabytes.

Extract, Transform and Load (ETL) – a process in a database and data warehousing meaning extracting the data from various sources, transforming it to fit operational needs and loading it into the database

Failover – switching automatically to a different server or node should one fail

Fault-tolerant design – a system designed to continue working even if certain parts fail Feature - a piece of measurable information about something, for example features you might store about a set of people, are age, gender and income.

Graph Databases – they use graph structures (a finite set of ordered pairs or certain entities), with edges, properties and nodes for data storage. It provides index-free adjacency, meaning that every element is directly linked to its neighbour element.

Grid computing – connecting different computer systems from various location, often via the cloud, to reach a common goal

Hadoop – an open-source framework that is built to enable the process and storage of big data across a distributed file system

HBase – an open source, non-relational, distributed database running in conjunction with Hadoop

HDFS – Hadoop Distributed File System; a distributed file system designed to run on commodity hardware

High-Performance-Computing (HPC) – using supercomputers to solve highly complex and advanced computing problems

Histogram - A graphical representation of the distribution of a set of numeric data, usually a vertical bar graph

In-memory – a database management system stores data on the main memory instead of the disk, resulting is very fast processing, storing and loading of the data

JavaScript - a scripting language designed in the mid-1990s for embedding logic in web pages, but which later evolved into a more general-purpose development language.

Juridical data compliance – relevant when you use cloud solutions and where the data is stored in a different country or continent. Be aware that data stored in a different country has to oblige to the law in that country.

KeyValue Databases – they store data with a primary key, a uniquely identifiable record, which makes easy and fast to look up. The data stored in a KeyValue is normally some kind of primitive of the programming language.

Latency – a measure of time delayed in a system

Load balancing – distributing workload across multiple computers or servers in order to achieve optimal results and utilization of the system

Location data – GPS data describing a geographical location

Log file – a file automatically created by a computer to record events that occur while operational

Machine data – data created by machines via sensors or algorithms

Machine learning – part of artificial intelligence where machines learn from what they are doing and become better over time

Metadata – data about data; gives information about what the data is about.

Multi-Dimensional Databases – a database optimized for data online analytical processing (OLAP) applications and for data warehousing.

MultiValue Databases– they are a type of NoSQL and multidimensional databases that understand 3 dimensional data directly. They are primarily giant strings that are perfect for manipulating HTML and XML strings directly

Natural Language Processing– a field of computer science involved with interactions between computers and human languages

Network analysis– viewing relationships among the nodes in terms of the network or graph theory, meaning analysing connections between nodes in a network and the strength of the ties.

Neural network – A type of artificial intelligence that attempts to imitate the way a human brain works. Rather than using a digital model, in which all computations manipulate zeros and ones, a neural network works by creating connections between processing elements, the computer equivalent of neurons. The organization and weights of the connections determine the output.

Object Databases – they store data in the form of objects, as used by object-oriented programming. They are different from relational or graph databases and most of them offer a query language that allows object to be found with a declarative programming approach.

Object-based Image Analysis – analysing digital images can be performed with data from individual pixels, whereas object-based image analysis uses data from a selection of related pixels, called objects or image objects.

Operational Databases – they carry out regular operations of an organisation and are generally very important to a business. They generally use online transaction processing that allows them to enter, collect and retrieve specific information about the company.

Optimization analysis - the process of optimization during the design cycle of products done by algorithms. It allows companies to virtually design many different variations of a product and to test that product against pre-set variables.

Pattern Recognition – identifying patterns in data via algorithms to make predictions of new data coming from the same source.

Petabytes - approximately 1000 terabytes or 1 million gigabytes. The CERN Large Hydron Collider generates approximately 1 petabyte per second

Predictive analysis – analysis within big data to help predict how someone will behave in the (near) future. It uses a variety of different data sets such as historical, transactional, or social profile data to identify risks and opportunities.

Public data – public information or data sets that were created with public funding

Query – asking for information to answer a certain question

Re-identification – combining several data sets to find a certain person within anonymized data

Regression analysis – to define the dependency between variables. It assumes a one-way causal effect from one variable to the response of another variable.

Real-time data – data that is created, processed, stored, analysed and visualized within milliseconds

Scripting - the use of a computer language where your program, or script, can be run directly with no need to first compile it to binary code. Semi-structured data - a form a structured data that does not have a formal structure like structured data. It does however have tags or other markers to enforce hierarchy of records.

Similarity searches – finding the closest object to a query in a database, where the data object can be of any type of data.

Simulation analysis – a simulation is the imitation of the operation of a real-world process or system. A simulation analysis helps to ensure optimal product performance taking into account many different variables.

Smart grid – refers to using sensors within an energy grid to monitor what is going on in real-time helping to increase efficiency

Spatial analysis – refers to analysing spatial data such geographic data or topological data to identify and understand patterns and regularities within data distributed in geographic space.

SQL – a programming language for retrieving data from a relational database

Structured data – data that is identifiable as it is organized in structure like rows and columns.

Terabytes – approximately 1000 gigabytes.

Time series analysis - analysing well-defined data obtained through repeated measurements of time. The data has to be well defined and measured at successive points in time spaced at identical time intervals.

Topological Data Analysis – focusing on the shape of complex data and identifying clusters and any statistical significance that is present within that data.

Un-structured data - unstructured data is regarded as data that is in general text heavy, but may also contain dates, numbers and facts.

Variability – it means that the meaning of the data can change (rapidly). In (almost) the same tweets for example a word can have a totally different meaning

Variety – data today comes in many different formats: structured data, semi-structured data, unstructured data and even complex structured data

Velocity – the speed at which the data is created, stored, analysed and visualized

Veracity – ensuring that the data is correct as well as the analyses performed on the data are correct.

Visualization – complex graphs that can include many variables of data while still remaining understandable and readable

Volume – the amount of data, ranging from megabytes to brontobytes

XML Databases – XML Databases allow data to be stored in XML format. The data stored in an XML database can be queried, exported and serialized into any format needed.

Yottabytes – approximately 1000 Zettabytes, or 250 trillion DVD's. The entire digital universe today is 1 Yottabyte and this will double every 18 months.

Zettabytes – approximately 1000 Exabytes or 1 billion terabytes