Investigating multi-species patterns across Fraser River salmon populations

Michael Barrus (michaelbarrus@psych.ubc.ca)

Introduction

Fisheries managers for the federal Department of Fisheries and Oceans are tasked with determining when and where to conduct fisheries, as well as how many fish to harvest at a time. These decisions are informed by advice from fisheries scientists, who in turn are responsible for understanding the factors that affect fish population dynamics. The salmon fishery is the most economically and culturally important fishery in British Columbia, and it is also perhaps the most complex. BC's salmon population is composed of hundreds of geographically and biologically distinct sub-populations. Salmon fisheries typically harvest multiple populations, and often even multiple species of salmon. These sub-groups exhibit divergent trends across characteristics such as productivity, abundance, and/or run timing, and this complexity can make management decisions difficult.

Understanding both local and global salmon population trends and the factors that drive them is essential to effective and sustainable fisheries management. However, there is presently no interface with which scientists and managers can review and examine existing salmon data across species and watersheds. The development of such an interface would provide salmon scientists and managers with a more thorough understanding of the state of salmon in British Columbia, as well as a stronger foundation for decision making given the challenging trade-offs under uncertainty that are inherent in multi-stock and multi-species fisheries.

This project will consolidate available data from Fraser salmon populations and provide a simple, interactive interface through which managers and researchers can review that data, identify patterns, examine how these patterns change over time, explore possible causes of trends, and evaluate their management options. This work will be guided by a consultation process which engages end users throughout the design process, in order to guarantee that the product addresses the specific, existing needs of the managers and researchers who work within the Fraser watershed.

In sum, this project has the following aims:

1) Develop a suite of visualizations that enable researchers within the department to examine the existing Fraser salmon population data
2) Develop a methodology with which to evaluate the ability of this vis to promote insights and hypothesis generation about factors affecting salmon populations
3) Develop a design plan for future visualizations of this data that exceed the scope of the course.

The development of this tool will provide salmon scientists with a stronger foundation for forecasting salmon abundances and fisheries expectations, and for estimating fisheries reference points. It will also provide salmon managers and stakeholders with additional factors to consider

in the challenging trade-offs under uncertainty that are inherent in multi-stock and multi-species fisheries.

I was recruited during summer 2017 by biologists within the DFO to work with them in developing such an interface after using a DFO dataset for a Tableau-based design class and posting it to Tableau public. My expertise is limited to the last ~6 months of work with salmon datasets and ~3 months of discussion with the DFO regarding the present dataset. Work performed prior to the start of the course has been limited to data cleaning, and applications for financial support; there has been little technical implementation to date. The work proposed within this document is novel, and is being treated as a springboard into a larger, full –scale design study slated to wrap up in spring 2019.

Data and tasks

The Fraser River Salmon dataset that will serve as the foundation for this project contains approximately seven categorical variables that characterize the species, location of origin and migration patterns of the salmon populations. The number of values associated with each category range from 3 (species) to 40 (location of origin). There are another 15 ordinal variables that describe various measure of population health. All of these non-categorical variables are organized as a time series with annual values, beginning in 1950 and concluding in 2015.

My discussions with department scientists have centered on the use of vis as a tool to promote understanding of this complex dataset, hypothesis generation and novel insights. The inherent fuzziness of this need suggests that an essential component of this project should be the development of empirical methods with which to evaluate the ability of the visualization to promote understanding. In addition to this fuzzy task, department scientists have described some more crisply defined ones such as recognizing patterns, finding emerging trends, clustering populations based on commonalities, and detecting extrema, correlations and anomalies.

Infovis solution

Methodology design. One merit of well-designed information visualization is that it allows users to rapidly explore and understand large datasets, and develop ideas and insights about those datasets. However, the concepts of ‘understanding’, ‘ideas’ and ‘insights’ are somewhat vague, and it is not immediately obvious how to quantify a visualization’s ability to promote those things. Given that the proposed visualization is being designed in part to encourage hypothesis generation, as well as ‘understanding’, ‘ideas’ and ‘insights’, it seems essential to develop a framework to evaluate how well our visualization accomplishes these goals. Having quantitative measures of these nebulous concepts will help us determine how well our tool is performing, and provides information about our accomplishments and progress to any external entities that have an interest in our work-- be they funding agencies, colleagues, higher-ups, etc. This methodology will be based on existing work such as [1], which similarly attempts to gauge the ability of visualization to improve people’s understanding of complex datasets. This methodology will be designed with the intention of being idiom-agnostic, so it can be applied to any visualizations generated as part of this project.

Tableau dashboard. The primary visualization that will be developed will be a Tableau dashboard, with both temporal and geographical representations of the dataset. Important features will be filters for categorical variables, small multiples, and the ability to perform similarity clustering across the time series. Figure 1 is the time-series encoding I've worked up thus far, while figure 2 is a geographic encoding that I constructed for a previous project; I expect that the final dashboard will be similar to both. The imagined use scenario would be exploratory in nature, much like the user testing described in [1]. Users would be free to use filters to select categorically-defined subsets of the dataset (such as species/location of origin/range of years/adaptive zone of interest), and examine these for similarities and differences.

Sandboxing future visualizations.

In addition to developing a methodology and a Tableau dashboard, I am interested in developing some visual encodings that may be beyond the scope of this course. Essentially, my ambition exceeds my present skills, and while there are visual encodings I think would be useful for exploring this data, I'm not capable of programming them. Rather than promising to have complete encodings done by the end of the course, I intend to dedicate some time to exploring the development of these idioms and beginning to work with existing libraries in order to do so. One of these is described below, while two more are listed but not described as they are existing idioms.

Quasi-geo space. Department scientists have expressed interest in understanding the geographic relationships between stocks, as representing the stocks' physical locations enables them to think about possible commonalities in environmental factors that can affect the health of these populations. Events such as landslides, wildfires, industrial activity and industrial accidents (such as the Mount Polley Mine tailings spill) may have effects on stocks both in the immediate vicinity of the incident as well as downstream. Plotting information about stock health in a way that conserves the physical relationships between the different populations would make it easier for researchers to detect patterns that are driven by shared geography. Furthermore, exactly replicating the geography of the region is not necessary to show these relationships and might in some cases obscure them. Plotting the data in a 'quasi-geographic' manner in which the relationships between systems is conserved while less-relevant information is filtered out may be a more effective approach than true geographic representation. I'm interested in developing an idiom that combines heat map small multiples (which encode population measures over time by river) with a dendrogram-like structure that represents the relationships between river systems which feed into each other, while obscuring the extraneous geographic information. A mock-up of this encoding is included as figure 3.

Cluster heat maps [2]

Data stripes [3]

Timeline

Category	Task	Req. time	Deadline	Description
Coursework	Pitch	4	Oct 17th	
Coursework	Proposal	8	Nov 6th	
Coursework	Peer review 1	2	Nov 21st	
Coursework	Peer review 2	2	Dec 6th	
Coursework	Final presentation	5	Dec 12th	
Coursework	Final paper	10	Dec 15th	
Design	Interview w/ end users	3	Nov 10th	
Design	Data cleaning	5	Nov 13-15th	
Design	Methodology write up	5	Nov 13th-15th	
Design	Tableau iteration 1	10	Nov 30th	
Design	User testing	3	Dec 1st	
Design	Tableau iteration 2	10	Dec 10th	
Design	Sandboxing	15	Dec 10th	Quasi- geo space
Design	Sandboxing	10	Dec 10th	Heatmap clustering
Design	Sandboxing	10	Dec 10th	Data stripes
Total		102		

Bibliography

[1] P. Saraiya, C. North and K. Duca, "An insight-based methodology for evaluating bioinformatics visualizations," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 4, pp. 443-456, July-Aug. 2005.

[2] L. Wilkinson and M. Friendly. "The History of the Cluster Heat Map," in *The American Statistician*, Vol. 63, No. 2, pp 179-184, May 2009

[3] Carl Manaster. "Data Stripes". https://github.com/carlmanaster/datastripes

Figure 1:

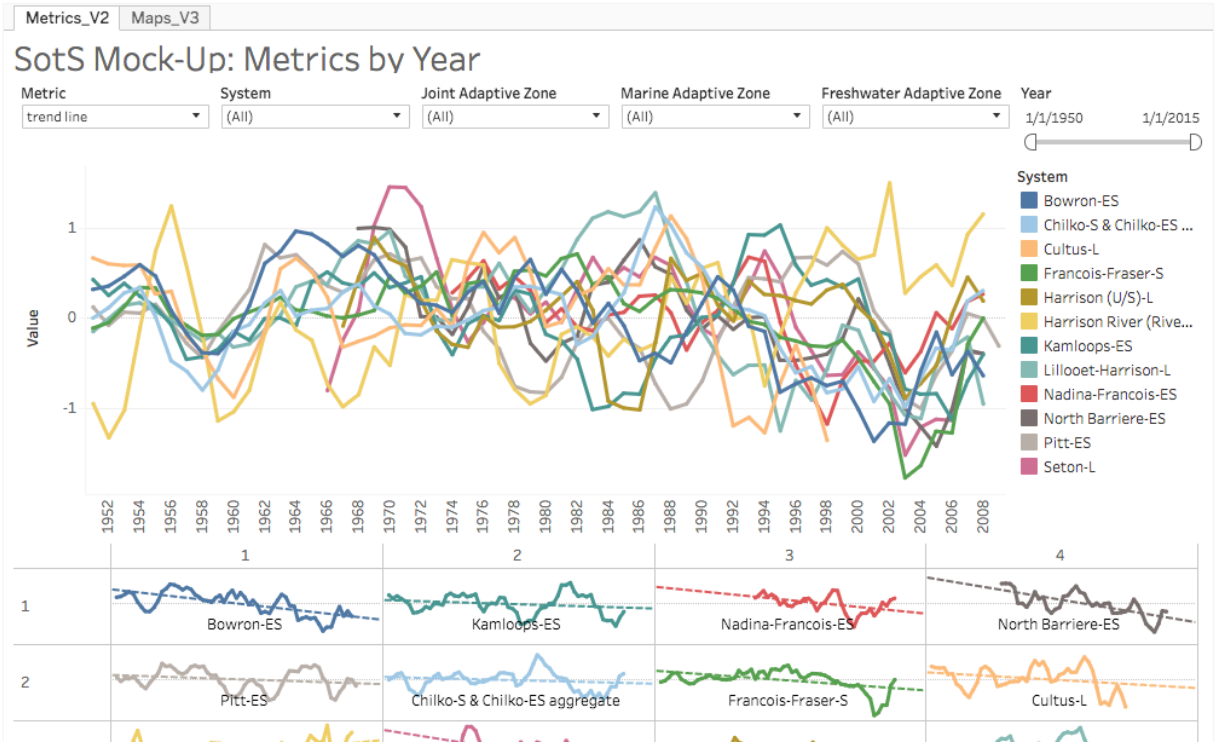
Metrics_V2
Maps_V3
SotS Mock-Up: Metrics by Year
Metric
trend line
System
(All)
Joint Adaptive Zone
(All)
Marine Adaptive Zone
(All)
Freshwater Adaptive Zone
(All)
Year
1/1/1950
1/1/2015
Value
1
0
-1
1952
1954
1956
1958
1960
1962
1964
1966
1968
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
System
Bowron-ES
Chilko-S & Chilko-ES ...
Cultus-L
Francois-Fraser-S
Harrison (U/S)-L
Harrison River (Rive...
Kamloops-ES
Lillooet-Harrison-L
Nadina-Francois-ES
North Barriere-ES
Pitt-ES
Seton-L
1
2
3
4
Bowron-ES
Kamloops-ES
Nadina-Francois-ES
North Barriere-ES
Pitt-ES
Chilko-S & Chilko-ES aggregate
Francois-Fraser-S
Cultus-L

Figure 2:

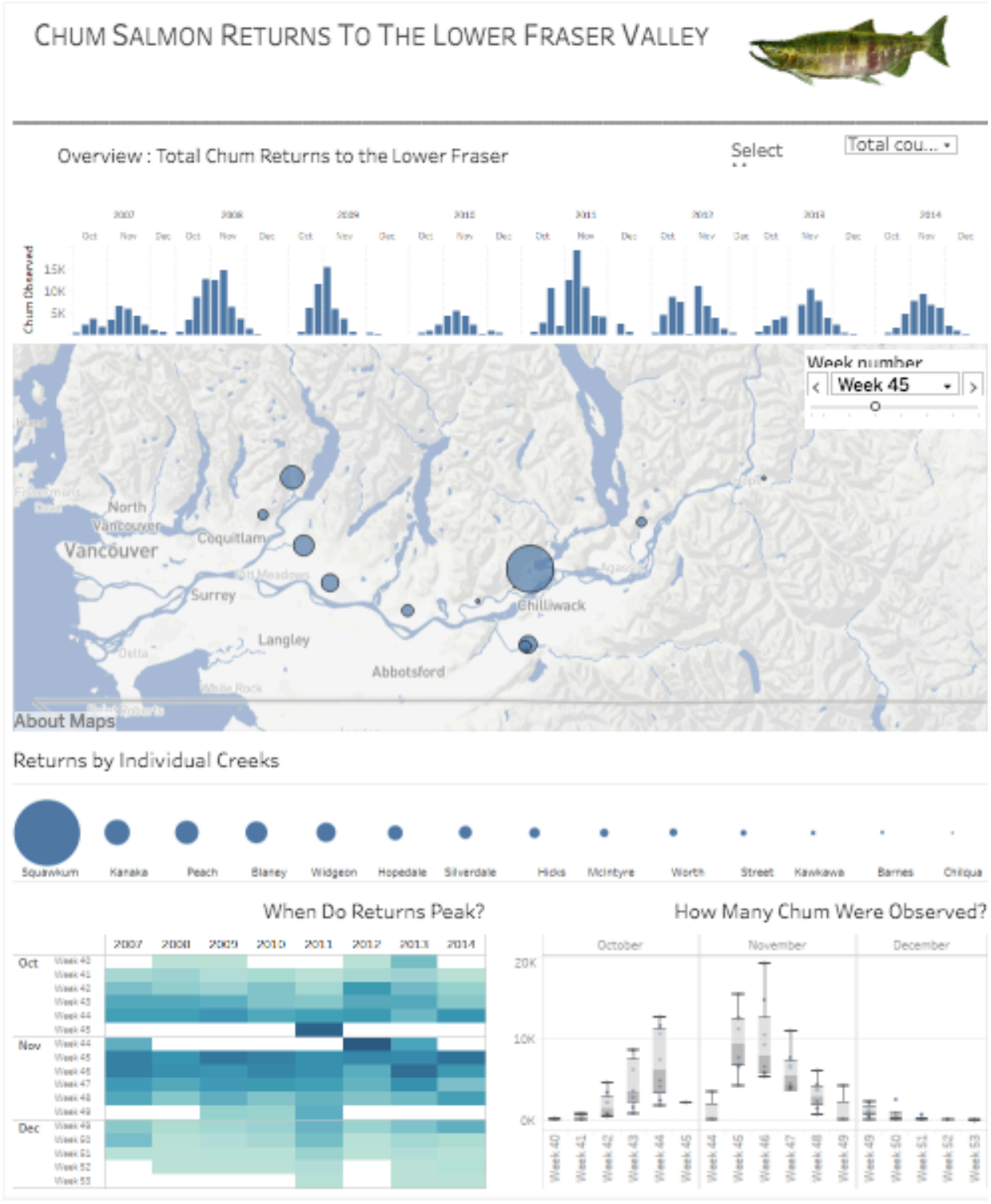
CHUM SALMON RETURNS TO THE LOWER FRASER VALLEY
Overview : Total Chum Returns to the Lower Fraser
Select
Total cou...
Chum Observed
15K
10K
5K
Week number
Week 45
North Vancouver
Vancouver
Coquitlam
Surrey
Chilliwack
Langley
Abbotsford
About Maps
Returns by Individual Creeks
Squawkum
Kanaka
Peach
Blaney
Widgeon
Hopedale
Silverdale
Hicks
McIntyre
Worth
Street
Kawkawa
Barnes
Chilqua
When Do Returns Peak?
How Many Chum Were Observed?
October
November
December
20K
10K
0K

Figure 3: