Contents

1 Survey of Quantitative Methods **2**
1.3 Data Visualisation/Representation 3
1.3.1 Pre-Analysis Use 3
1.3.2 Presenting Results 3
1.3.3 Multivariate Elements in Charts 4
1.3.4 A Word About Accessibility 9
1.3.5 Case Study: Reliability of Canadian Consular Network Data 14

List of Figures

1 Data visualisation suggestions, by type of question 5
2 Bubble Chart: Gapminder's Health and Wealth of Nations 6
3 Choropleth Map: mean elevation by U.S. state 7
4 Network Clustering: lexical distance of European languages 8
5 Word Cloud: U.S. Constitution and Canadian Charter of Rights and Liberties 9
6 Decision Tree: classification scheme for the kyphosis dataset 10
7 Histogram: artificial dataset 10
8 Clustering Scatterplot: clustering of Sandra Bullock movies 11
9 Decision Tree Bubble Chart: NASA's COCOMO dataset 11
10 Association Rule Network: Danish Medical Dataset 12
11 Classification Scatterplot: artificial dataset 12
12 Classification Bubble Chart: Hertzsprung-Russell Diagram 13
13 Time Series: trend, seasonality, and shifts 13

List of Tables

1 Survey of Quantitative Methods

The bread and butter of quantitative consulting is the ability to apply quantitative methods to business problems in order to obtain actionable insight. Clearly, it is impossible (and perhaps inadvisable, in a more general sense) for any given individual to have expertise in every field of mathematics, statistics, and computer science.

We believe that the best consulting framework is reached when a small team of consultants possesses expertise in 2 or 3 areas, as well as a decent understanding of related disciplines, and a passing knowledge in a variety of other domains: this includes keeping up with trends, implementing knowledge redundancies on the team, being conversant in non-expertise areas, and knowing where to find detailed information (online, in books, or through external resources).

In this section, we present an introduction for 9 "domains" of quantitative analysis:

- survey sampling and data collection;
- data processing;
- data visualisation;
- statistical methods;
- queueing models;
- data science and machine learning;
- simulations;
- optimisation, and
- trend extraction and forecasting;

Strictly speaking, the domains are not free of overlaps. Large swaths of data science and time series analysis methods are quite simply statistical in nature, and it's not unusual to view optimisation methods and queueing models as sub-disciplines of operations research. Other topics could also have been included (such as Bayesian data analysis or signal processing, to name but two), and might find their way into a second edition of this book.

Our treatment of these topics, by design, is brief and incomplete. Each module is directed at students who have a background in other quantitative methods, but not necessarily in the topic under consideration. Our goal is to provide a quick "reference map" of the field, together with a general idea of its challenges and common traps, in order to highlight opportunities for application in a consulting context. These subsections are emphatically NOT meant as comprehensive surveys: they focus on the basics and talking points; perhaps more importantly, a copious number of references are also provided.

We will start by introducing a number of motivating problems, which, for the most part, we have encountered in our own practices. Some of these examples are reported on in more details in subsequent sections, accompanied with (partial) deliverables in the form of charts, case study write-ups, report extract, etc.).

As a final note, we would like to stress the following: it is **IMPERATIVE** that quantitative consultants remember that acceptable business solutions are not always optimal theoretical solutions. Rigour, while encouraged, often must take a backseat to applicability. This lesson can be difficult to accept, and has been the downfall of many a promising candidate.

1.3 Data Visualisation/Representation

Analysis in the Modern Age

Discovery is no longer limited by the collection and processing of data, but rather management, analysis, and visualisation.

– @DamienMingle

What can be done with the data, once it has been collected? Two suggestions come to mind:

- **analysis** is the process by which we extract actionable insights from the data (this process is discussed in later subsections), while
- **visualisation** is the process of presenting data, calculations, and analysis outputs in a visual format. Visualisation of data *prior* to analysis can help simplify the analytical process. Visualisation *following* analysis allows for the analysis results to be presented to various stakeholders.

In this section, we focus on important visualisation concepts and methods; we shall provide examples of data displays to illustrate the various possibilities that might be produced by the data presentation component of a data analysis system.

1.3.1 Pre-Analysis Use

The Ying..

A picture is worth a thousand words.

– ancient saying

Even before the analytical stage is reached, data visualisation can be used to set the stage for analysis by:

- detecting invalid entries and outliers,
- shaping the data transformations (binning, standardisation, Box-Cox transformations, dimension reduction, etc.),
- getting a sense for the data (data analysis as an art form, exploratory analysis), and
- identifying hidden data structures (clustering, associations, patterns which may inform the next stage of analysis, etc.)

1.3.2 Presenting Results

... and the Yang

A thousand and one words are worth more than a picture.

– John McCarthy (attributed)

The crucial element of data presentations is that they need to help convey the insight or the message. To that effect, they should be clear, engaging, and (more importantly) readable. Our ability

to think of questions (and to answer them) is in some sense limited by what we can visualise. There is always a danger that if certain types of visualisation techniques dominate the evidence presentations, the kinds of questions that are particularly well-suited to providing data for these techniques will come to dominate the landscape, which will then affect data collection techniques, data availability, future interest, and so forth.

Generating Ideas and Insights In *Beautiful Evidence* [9], Edward Tufte explains that evidence is presented to assist our thinking processes. He further suggests that there is a symmetry to visual displays of evidence – the consumers should be seeking exactly what the producers should be providing, namely:

- meaningful comparisons
- causal networks and underlying structure
- multivariate links
- integrated and relevant data
- a primary focus on content

The choice of visualisation methods is strongly dependent on the analysis objective, that is, on the questions that need to be answered. The presentation method should not be selected randomly (or simply from a list of easily-produced templates).

In Figure 1, Frédérik Ruys suggests various types of visual displays that can be used, depending on the questions that are being asked:

- who is involved?
- where is it taking place?
- when did it happen?
- what is it about?
- how/why does it work?
- how much?

A general dashboard should at least be able to produce the following types of display:

- **charts** – comparison and relation (scatterplots, bubble charts, parallel coordinate charts, decision trees, cluster plots, trend plots)
- **choropleth maps** (heat maps, classification maps)
- **network diagrams** and connection maps (association rule networks, phrase nets)
- **univariate diagrams** (word clouds, box plots, histograms)

1.3.3 Multivariate Elements in Charts

Cubism's Missing Link

Picasso was particularly struck by Poincaré's advice on how to view the fourth dimension, which artists considered another spatial dimension. If you could transport yourself into it, you would see every perspective of a scene at once. But how to project these perspectives on to canvas?

– John McCarthy (attributed)

At most two fields can be represented by position in the plane. How can we then represent other crucial elements on a flat computer screen?

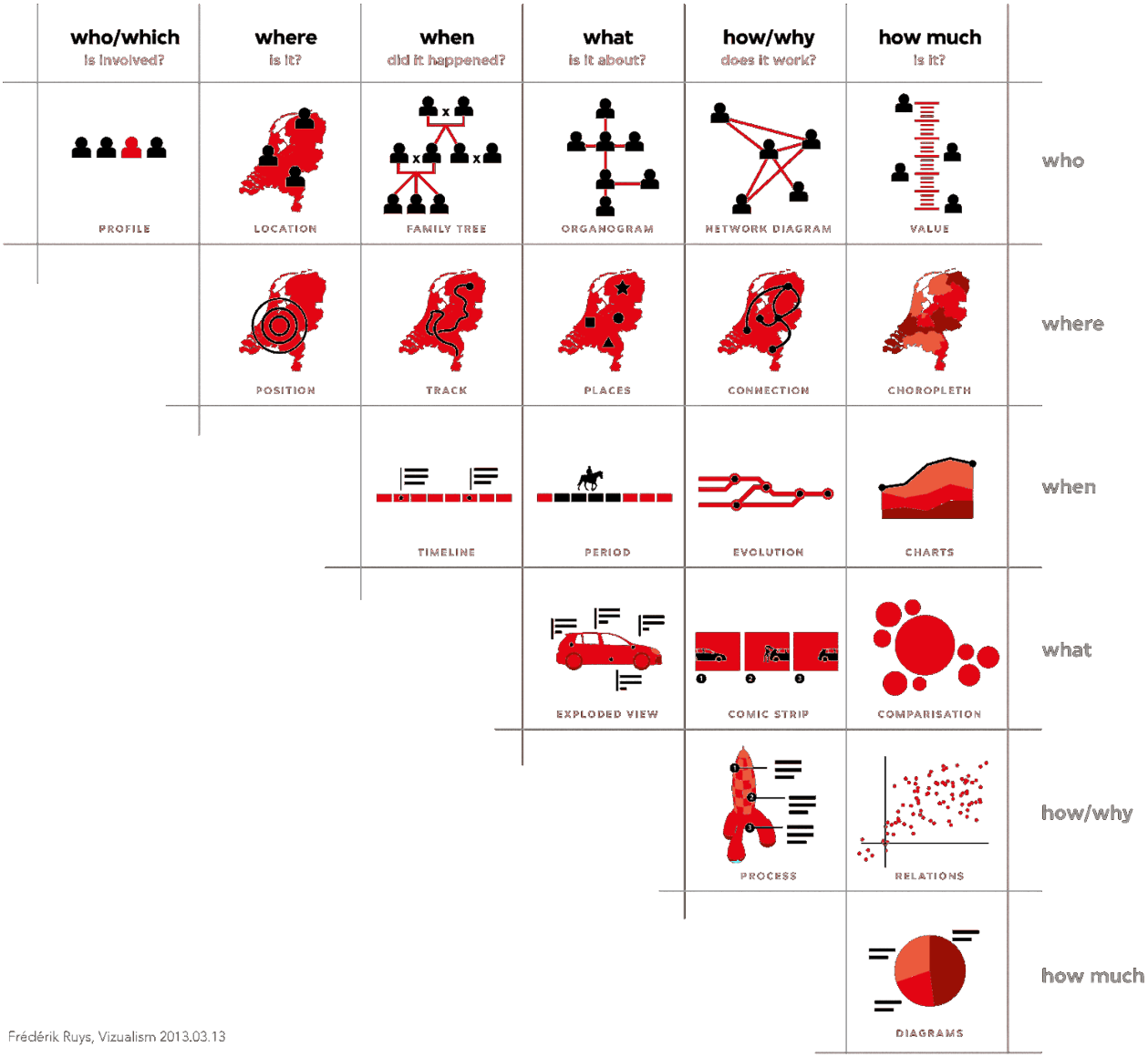


Figure 1: Data visualisation suggestions, by type of question (F.Ruys, Vizualism).

Potential solutions include:

- third dimension
- marker size
- marker colour
- colour intensity and value
- marker texture
- line orientation
- marker shape
- motion/movie

These elements do not always mix well – efficient design is as much art as it is science.

Examples In what follows, we provide examples for each of these chart types, along with a concise description of key components and a list of questions that they could help answer for four of them. Some additional diagrams showcasing the four presentation types discussed previously are highlighted in Figures 6 to 13.

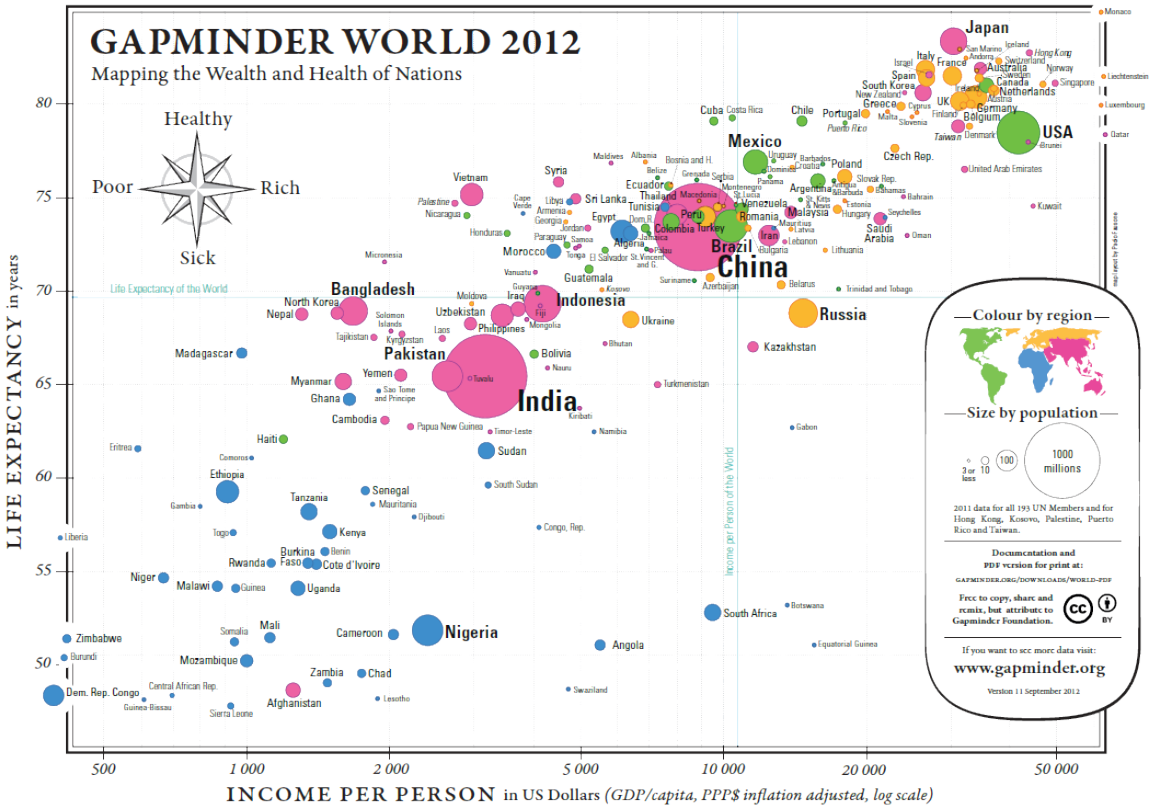


Figure 2: Gapminder's Health and Wealth of Nations (H.Rosling).

Bubble Chart: Health and Wealth of Nations (Figure 2)

- **Data:** 2011 life expectancy in years, inflation adjusted GDP/capita in USD, population for 193 UN members and 5 other countries.
- **Some Questions and Comparisons:** Can we predict the life expectancy of a nation given its GDP/capita? (*The trend is roughly linear: Expectancy* $\approx 6.8 \times \ln GDP/capita + 10.6$)
  Are there outlier countries? (*South Africa, Botswana, and Vietnam, at a glance*)
  Are countries with a smaller population healthier? (*Bubble size seems uncorrelated with the axes' variates*)
  Is continental membership an indicator of health and wealth levels? (*There is a clear divide between Western Nations (and Japan), most of Asia, and Africa*)
  How do countries compare against world values for life expectancy and GDP per capita? (*The vast majority of countries fall in three of the quadrants – there are very few wealthy countries with low life expectancy. China sits near the world values, which is expected for life expectancy, but more surprising when it comes to GDP/capita – compare with India*)
- **Multivariate Elements:** Scatterplot positions for health and wealth, bubble size for population, colour for continental membership, labels to identify the nations.
- **Comments:** Are life expectancy and GDP/capita appropriate proxies for health and wealth? A fifth element could also be added to a screen display: the passage of time.
- **Reference:** Image and documentation available at the Gapminder Foundation https://www.gapminder.org/downloads/world-pdf/

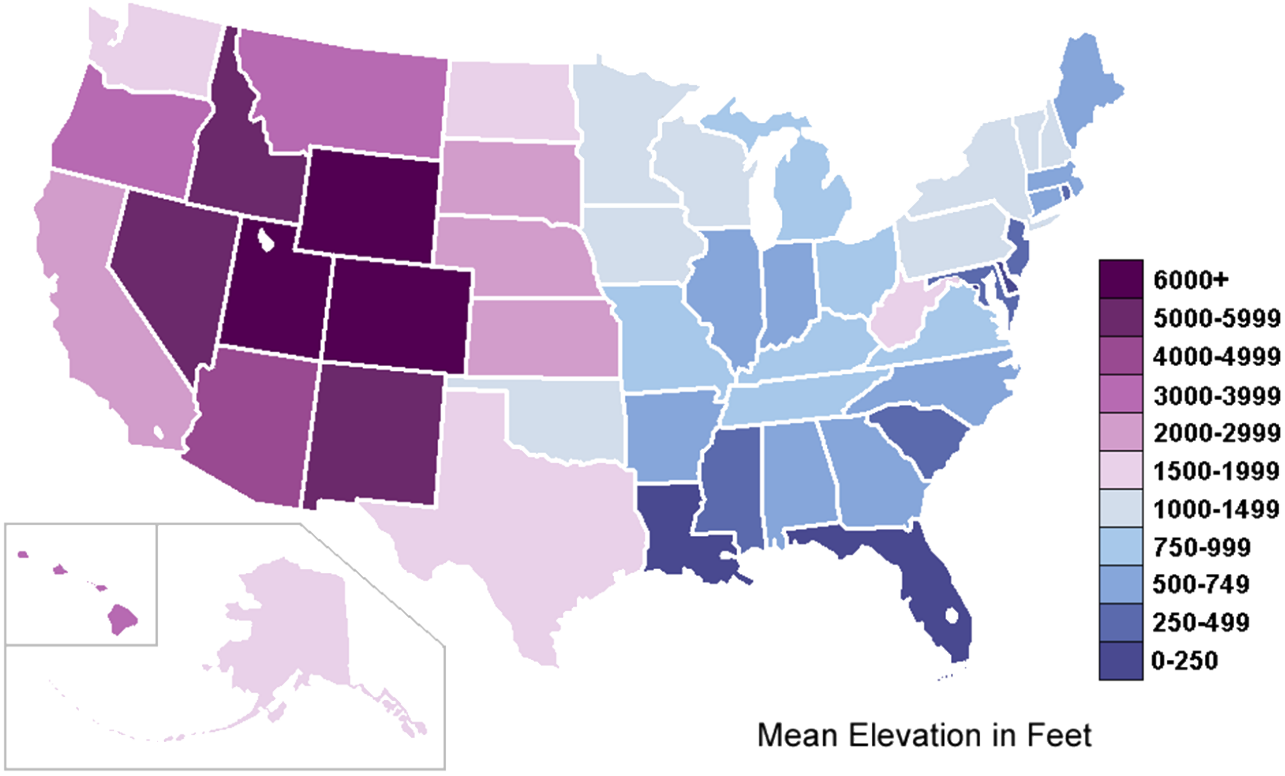


Figure 3: Mean elevation by U.S. state, in feet (source unknown).

Choropleth Map: Mean Elevation by U.S. State, in feet (Figure 3)

- **Data:** 50 observations, ranging from sea level (0-250) to (6000+)
- **Some Questions and Comparisons:** Can the mean elevation of the U.S. states tell us something about the global topography of the U.S.? (*West has higher mean elevation related to the presence of the Rockies; Eastern coastal states are more likely to suffer from rising water levels, for instance*)
  Are there any states that do not "belong" in their local neighbourhood, elevation-wise? (*West Virginia and Oklahoma seem to have the "wrong" shade – is that an artifact of the colour gradient and scale in use?*)
- **Multivariate Elements:** Geographical distribution and purple-blue colour gradient (as the marker for mean elevation)
- **Comments:** Is the 'mean' the right measurement to use for this map? (*it depends on the author's purpose.*)
  Would there be ways to include other variables in this chart? (*population density with texture, for instance*)
- **Reference:** Author unknown.

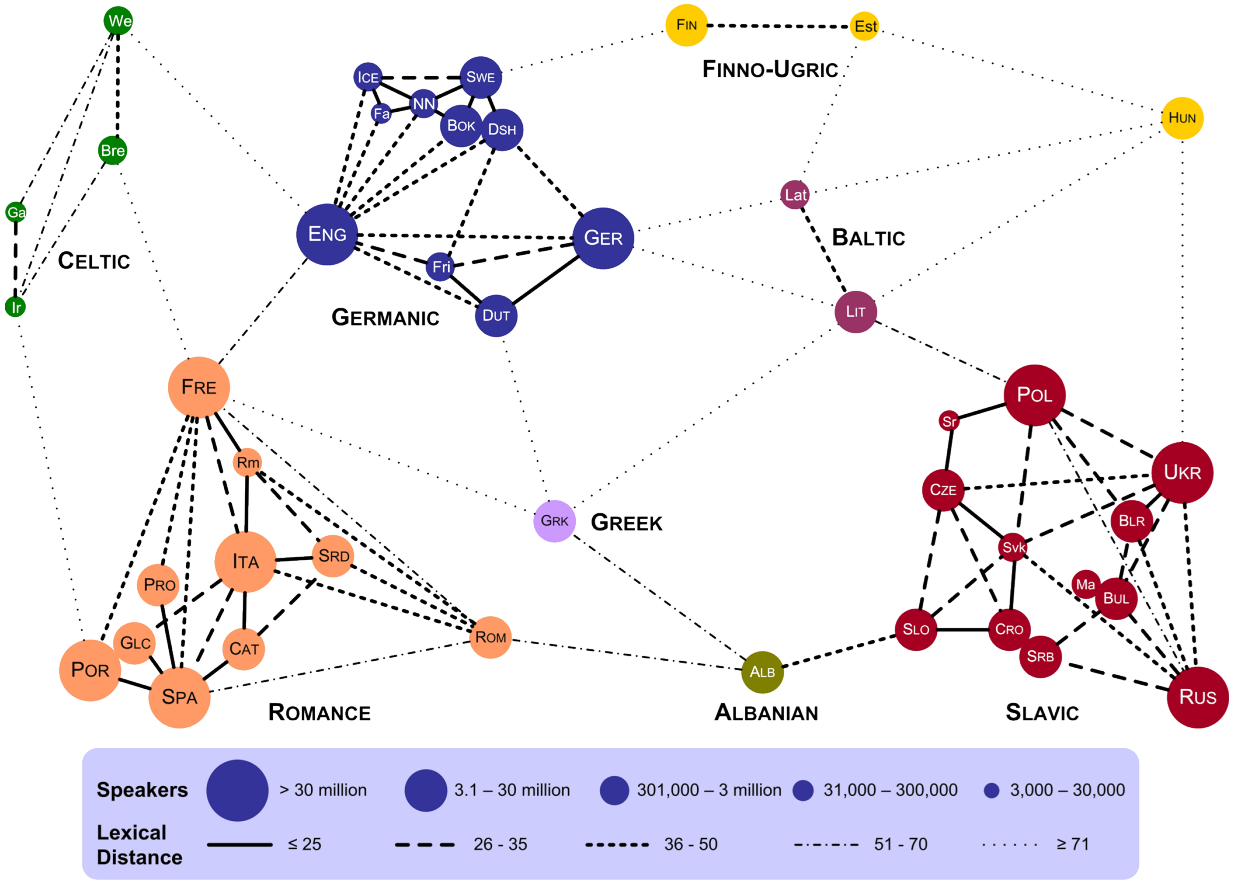


Figure 4: Lexical distance of European languages (T.Elms).

Network Diagram: Lexical Distance of European Languages (Figure 4)

- **Data:** Speakers and language groups for 43 European languages, lexical distances
- **Some Questions and Comparisons:** Are there languages that are lexically closer to languages in other lexical groups than to languages in their own groups? (*French is lexically closer to English than it is to Romanian*)
  Which language has the most links to other languages? *(English has 10 links)*
  Are there languages that are lexically close to multiple languages in other groups? (*Greek is lexically close to French (Romance), Albanian, Dutch (Germanic), and Lithuanian (Baltic)*)
  Is there a correlation between the number of speakers and the number of languages in a language group? *(Language groups with more speakers tend to have more languages)*
- **Multivariate Elements:** Colour and cluster for language group, line style for lexical distance, bubble size for number of speakers
- **Comments:** How is lexical distance computed?
  Some language pairs are not joined by links – does this mean that their lexical distance is large enough not to be rendered?
  Are the actual geometrical distances meaningful? For instance, Estonian is closer to French in the chart than it is to Portuguese – is it also lexically closer?
- **Reference:** Teresa Elms, Etymologikon
  https://elms.wordpress.com/2008/03/04/lexical-distance-among-languages-of-europe/

Figure 5: U.S. Constitution and Canadian Charter of Rights and Liberties.

Word Cloud: U.S. Constitution and Canadian Charter of Rights and Liberties (Figure 5).

- **Data:** Text version of the U.S. Constitution and the Canadian Charter of Rights and Liberties
- **Some Questions and Comparisons:** Are these two documents of the same type?
  Do the documents have the same authors?
  Could they conceivably have been written in the same era?
  Can important differences between Canada and the U.S. be gleamed from the wordclouds?
- **Univariate Element:** Font size correlated with frequency count.
- **Comments:** Note the absence of common, content-free words (the, and, etc.).
  Colour, orientation, and word placement are not mapped to multivariate data elements in these charts, but these options are available.
  Semantic parsing and phrase nets could give us a better idea of the general "sentiment" underlying the documents – what groups of words tend to occur together?
  Once the main differences have been absorbed, removing the most frequent terms from the data and producing new wordclouds on the updated texts can allow for insights regarding the setting.
- **Reference:** Personal file.

1.3.4 A Word About Accessibility

Cubism's Missing Link

Hear now this, O foolish people, and without understanding; which have eyes, and see not; which have ears, and hear not.

– Jeremiah 5:21 (King James Bible)

While visual displays can help provide analysts with insight, some work remains to be done in regard to visual impairment. A table can be translated to Braille fairly easily, but short of describing the features and emerging structures in a visualisation, even the cleverest of graphs will only succeed in conveying relevant information to a small subset of the population. The onus remains on the analyst to not only produce clear and meaningful visualisations, but also to describe them and their features in a fashion that allows all to "see" the insights. One drawback is that in order for this description to be done properly, the analyst needs to have seen all the insights, which is not necessarily the case (if at all possible).

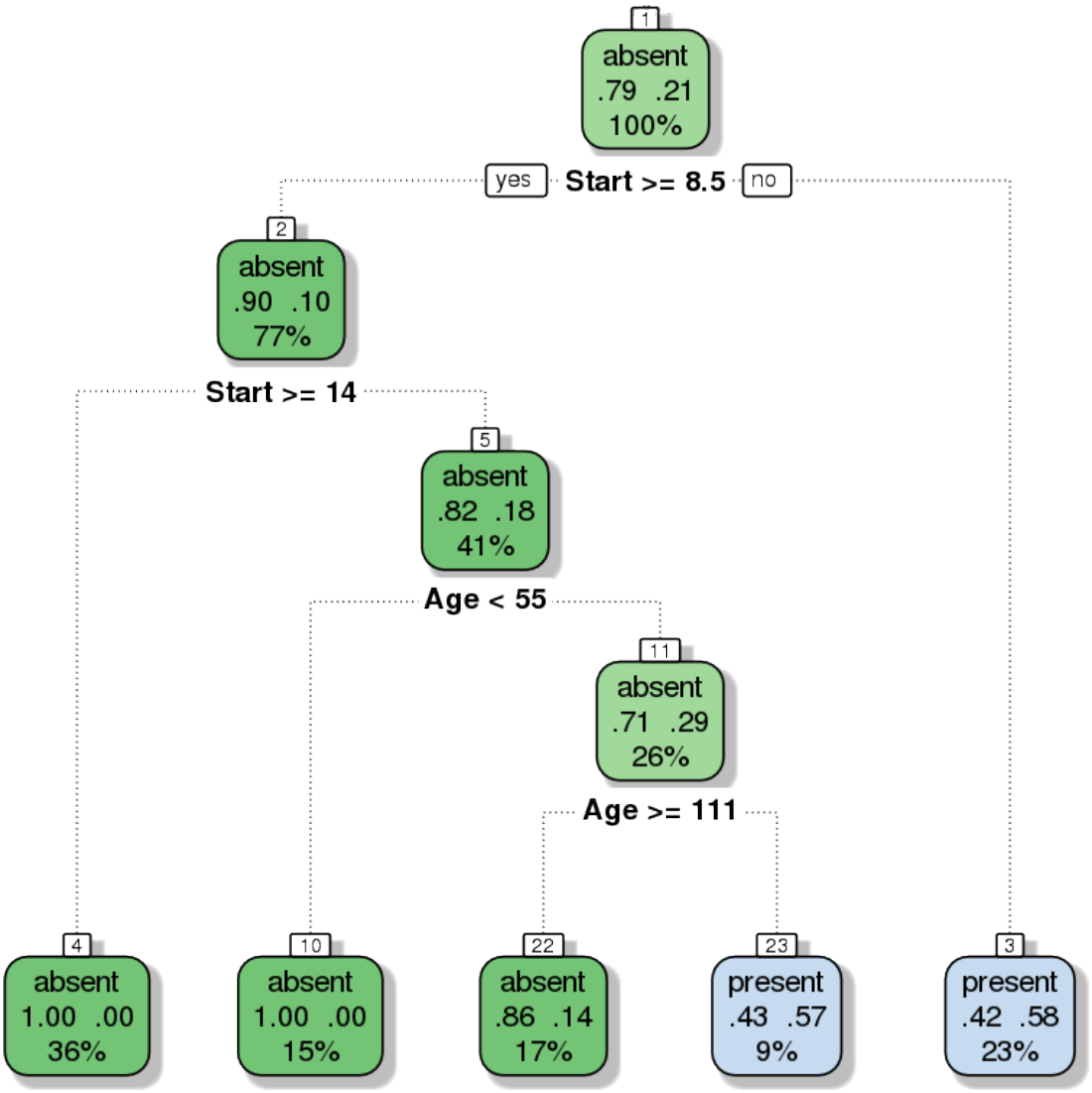


Figure 6: Decision Tree: classification scheme for the kyphosis dataset (personal file).

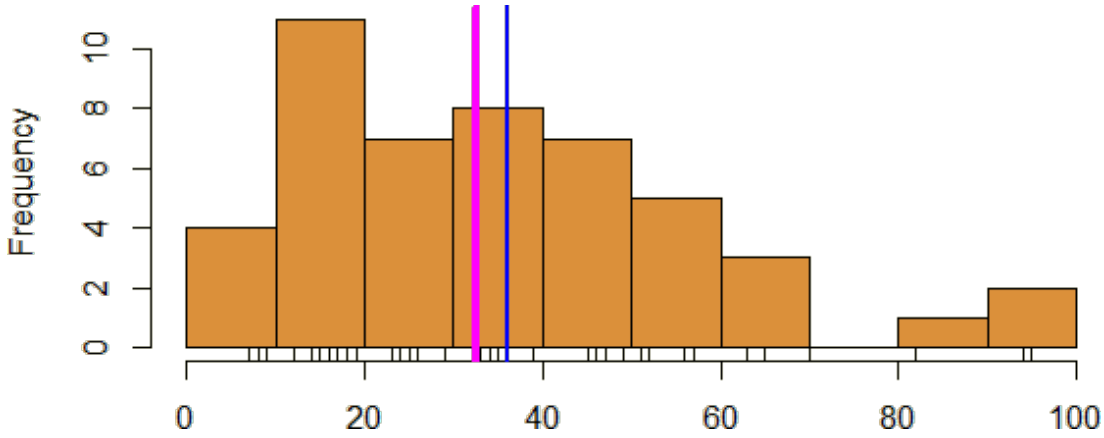


Figure 7: Histogram: artificial dataset (personal file).

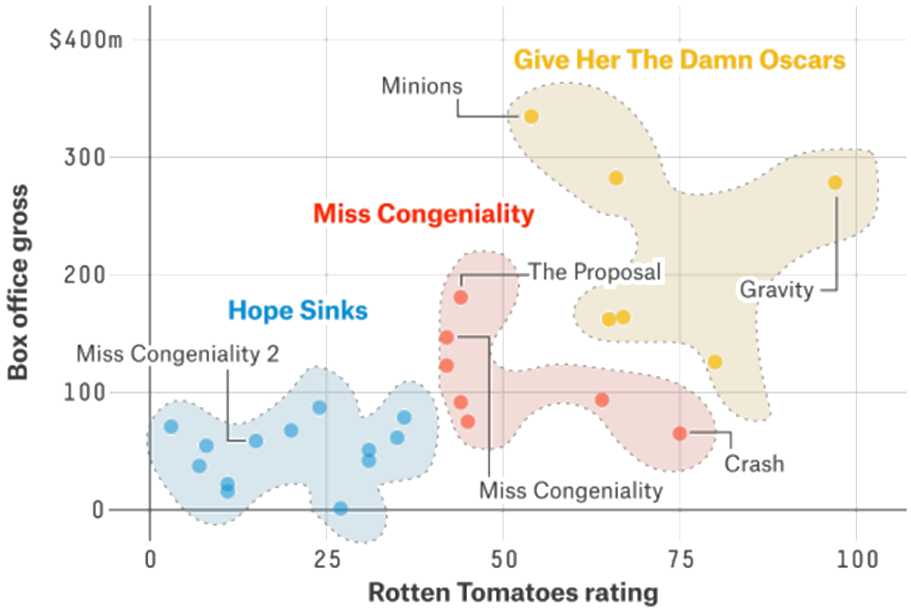


Figure 8: Clustering Scatterplot: clustering of Sandra Bullock movies, by Rotten Tomatoes rating and inflation-adusted domestic gross (fivethirtyeight.com).

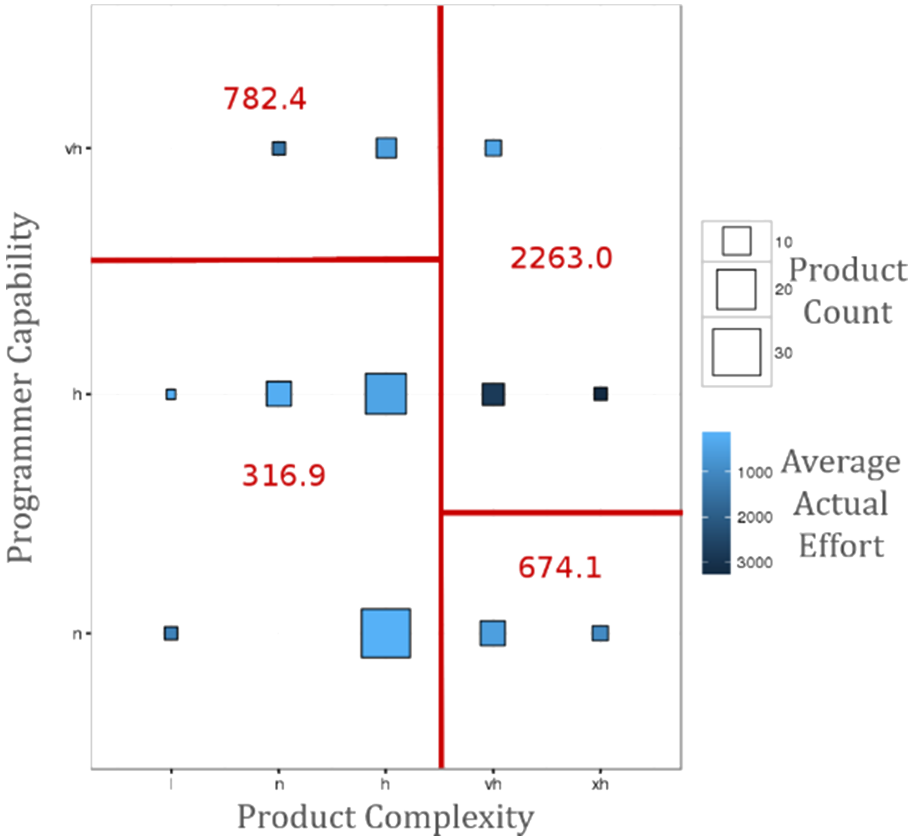


Figure 9: Decision Tree Bubble Chart: estimated average project effort (in red) over-layed over product complexity, programmer capability, and product count in NASA's COCOMO dataset (personal file).

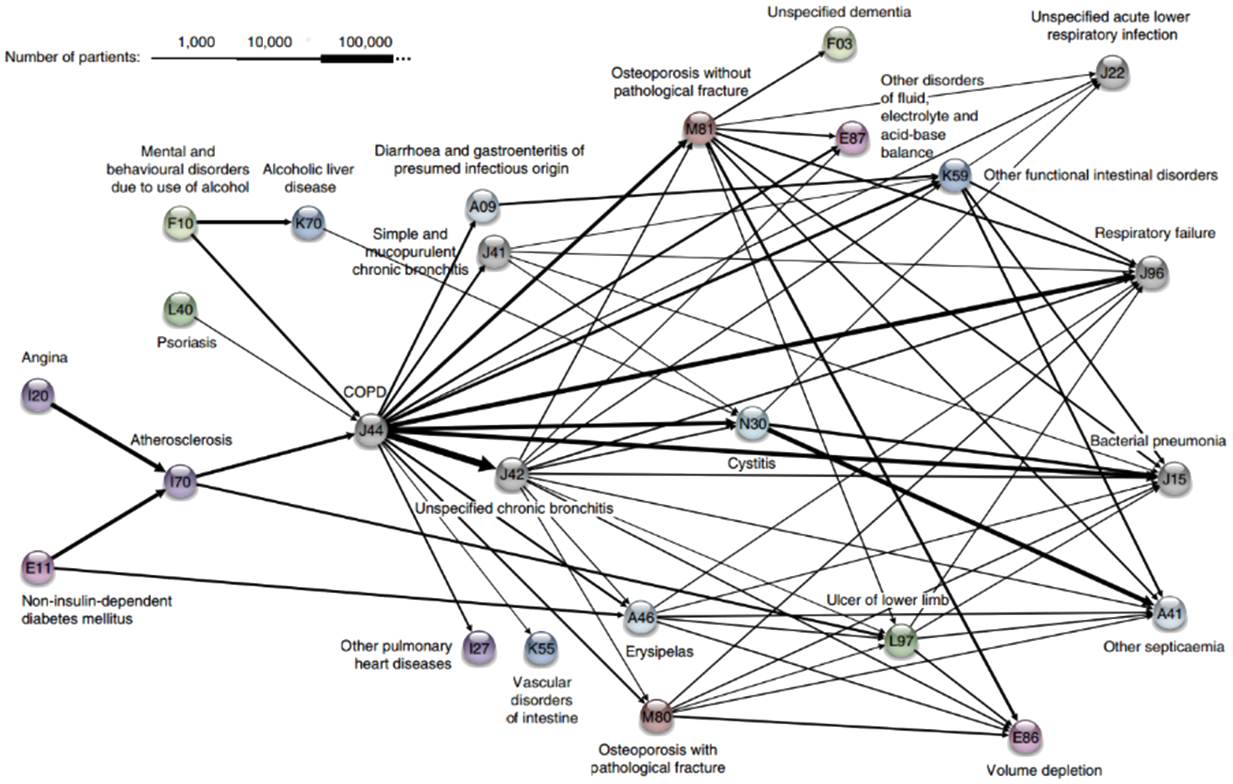


Figure 10: Association Rule Network: diagnosis network around COPD in the Danish Medical Dataset (A.B. Jensen *et al*).

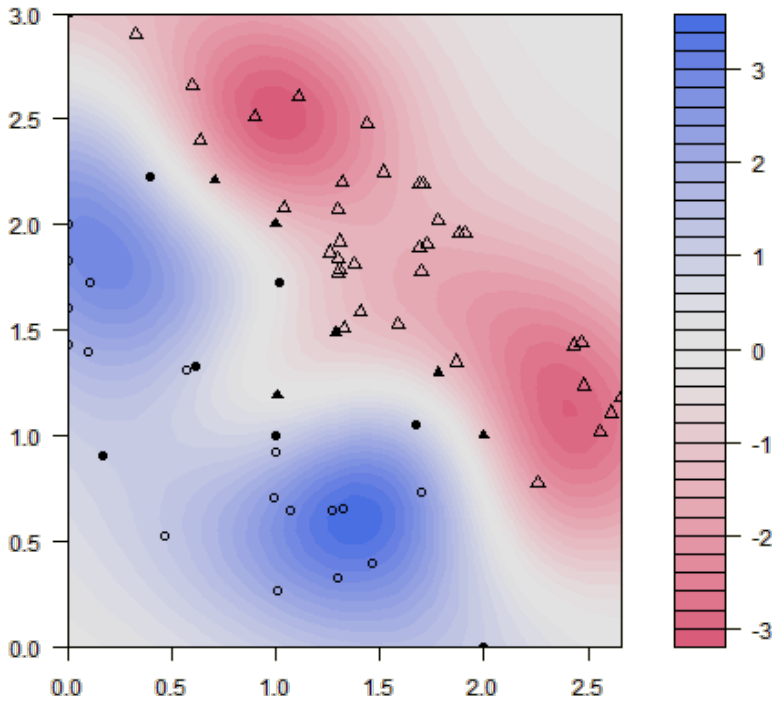


Figure 11: Classification Scatterplot: artificial dataset (personal file).

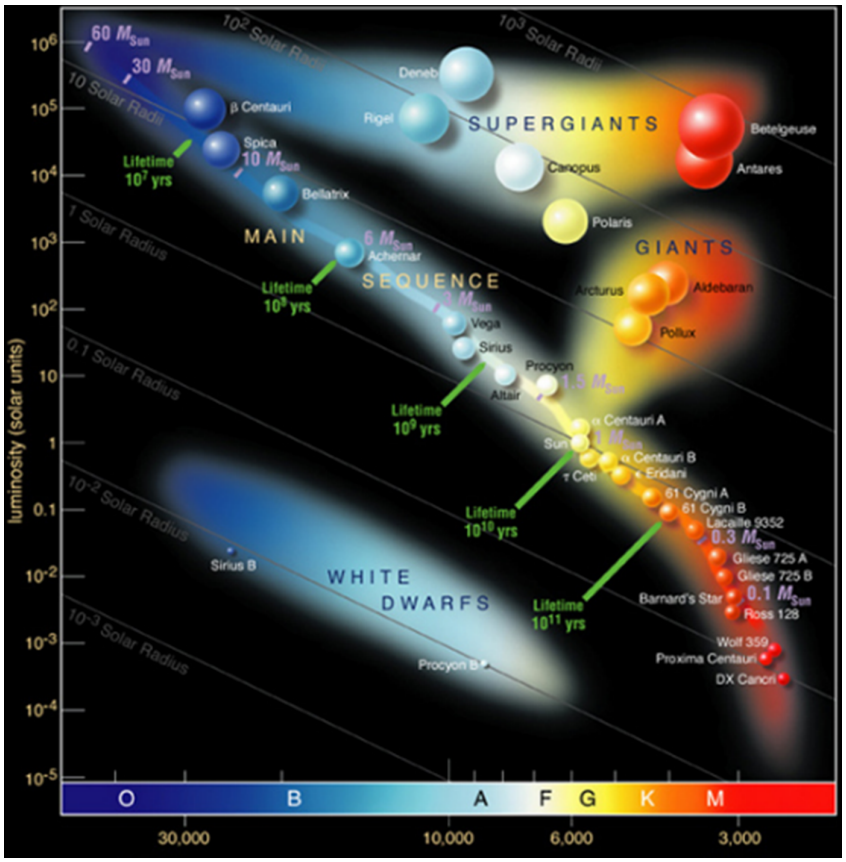


Figure 12: Classification Bubble Chart: Hertzsprung-Russell Diagram (European Southern Observatory).

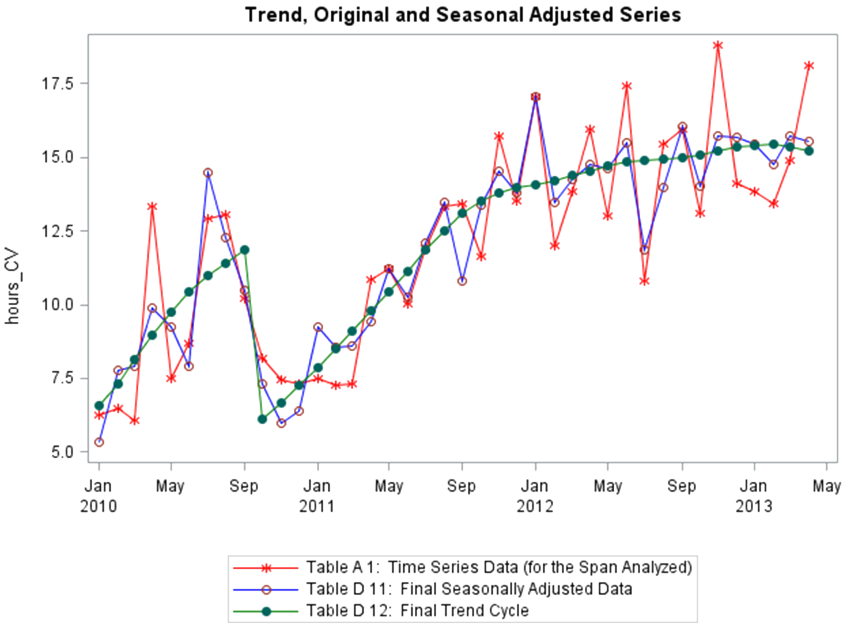


Figure 13: Time Series: trend, seasonality, and shifts (personal file).

1.3.5 Case Study: Reliability of Canadian Consular Network Data

The *Canadian Consular Network* provides services to Canadians travelling or living abroad (loss of a passport, need for urgent medical care, complications due to an arrest, or other emergencies). Consular officials can be reached 24 hours a day, seven days a week, at more than 260 points of service in 150 countries and through the *Emergency Watch and Response Centre* in Ottawa. The type and amount of help that consular officials can provide depends on the situation and may be affected by natural disasters, political unrest, and the laws in effect in other countries.

Within *Global Affairs Canada* (GAC), the Consular Corporate Management and Innovation group (CCMI) uses COSMOS, a software application that tracks consular activity statistics. COSMOS is used to enable consulates to provide assistance to their consular clients and to help identify where the workload stresses are located. It can also be used to provide basic statistics for requests from journalists and others.

COMIP (a COSMOS module) tracks the time required by employees to perform consular tasks. This data, in one form or another, stretches back over approximately twenty years (from 2016). It is currently used to determine the effectiveness of mission consular programs, to identify weaknesses to be resolved through HR, training and other solutions, and to evaluate resources need in missions – COMIP is the pivotal element when determining whether to staff, delete, or create positions. The software is scheduled to be updated/replaced (late 2016), and GAC would like to use this opportunity to determine if the current system meets their needs, and if not, how it could be improved.

The data of primary interest for consular management is contained in 4 COMIP tables: the logs of mission activities (**cases**, **services**, and **programs**), as well as the daily and monthly time spent on these mission activities, by employee. In these tables, data is available across a time span of 10 years, from 2005–2014; however, a 2010 system upgrade changed the categories relating to cases and services, resulting in a break in the dataset at this time.

In data analytical endeavours, the quality of the output is affected by the quality of the input, especially when it is self-reported (such as is the case with COMIP). CCMI understands that monthly log data has, in some sense, more inherent validity than daily log data as it must be reviewed by management before being submitted into the system – this oversight may be sufficient to ensure greater validity of that data. Daily log data by contrast may be entered less diligently as they are not required to produce a monthly log.

Abandoning daily log data altogether is not a solution as it is impossible to create monthly log data that accurately reflects the reality of monthly work in the mission without using (some) information from employees about their daily work during the month. Daily data, thus, is used *de facto* to create the monthly logs.

As a result, while certain types of consular data analysis may be conducted using monthly aggregates, data validation has to occur at the daily data level. In this case study, we present various data visualisations that were produced to study the reliability and validity of COMIP data.

Contents

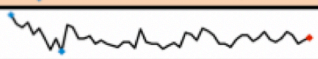
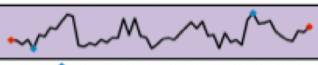
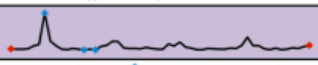
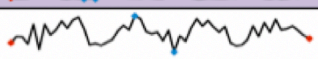
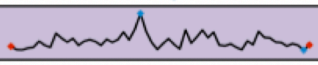
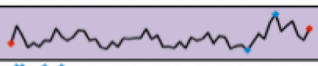
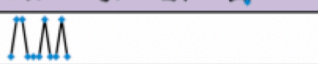
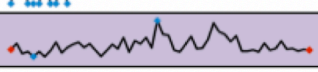
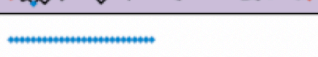
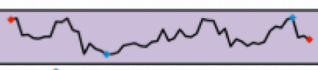
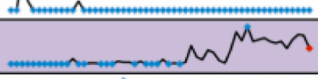
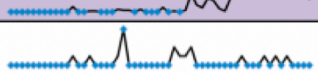
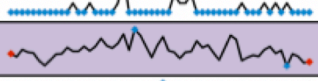
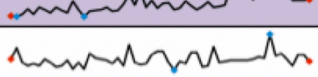
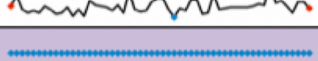
Basic Check - Data Gaps
Entire Dataset Review
Mission-Level Dataset Review
Plausibility of Work Hours
Employee-Level Dataset Review

References

[1] Malamed, C., **Understanding Graphics**.
[2] Krygier, J., Wood, D., [2016], Making Maps: A Visual Guide to Map Design for GIS, Guilford Press.
[3] **Interactive Data Visualization** on Wikipedia.
[4] Simmon, R. [2013], **Is animation an effective tool for data visualization?**, NASA's Earth Observatory.
[5] Healey, C.G., **Perception in Visualization**
[6] **Data Physicalizations**
[7] Tufte, E. [2001], The Visual Display of Quantitative Information, Graphics Press.
[8] Hu, D. [1954], How to Lie With Statistics, Norton.
[9] Tufte, E. [2008], Beautiful Evidence, Graphics Press.
[10] Nussbaumer Knaflic, C. [2015], Storytelling with Data, Wiley.
[11] Cairo, A. [2013], The Functional Art, New Riders.
[12] Cairo, A. [2016], The Truthful Art, New Riders.
[13] Meireilles, I. [2013], Design for Information, Rockport.
[14] **50 Great Examples of Data Visualization**.
[15] Kirk, A., **Visualising Data**.
[16] Yau, N., **FlowingData**.
[17] **Data Visualization** on Wikipedia.
[18] **Misleading Graphs** on Wikipedia.
[19] Prabhakaran, S., **Top 50 ggplot2 Visualizations**.
[20] Miller, M. [2017], The problem with Interactive graphics, Co.Design
[21] Wickham, H. [2016], ggplot2: Elegant Graphics for Data Analysis (2nd ed), Springer.
[22] Gorelik, B., Data Visualization (blog).
[23] Miller, A.I. [2012], **Henri Poincaré: the unlikely link between Einstein and Picasso**, in The Guardian.
[24] Wexler, S., Shaffer, J., Cotgreave, A. [2017], the Big Book of Dashboards, Wiley.

	1	2	3	4	5	6	7	8	9	10
1	66%	66%	58%	65%	66%	60%	14%	66%	13%	76%
2	1%	67%	60%	18%	50%	8%	66%	65%	66%	32%
3	17%	81%	53%	61%	46%	80%	65%	66%	61%	65%
4	73%	61%	63%	67%	67%	63%	44%	73%	66%	2%
5	66%	64%	63%	29%	35%	52%	71%	76%	66%	28%
6	64%	90%	57%	16%	1%	64%	64%	64%	7%	61%
7	3%	62%	59%	65%	31%	59%	64%	63%	67%	87%
8	11%	64%	66%	64%	38%	56%	65%	65%	65%	65%
9	78%	80%	69%	15%	26%	66%	66%	68%	63%	0%
10	73%	79%	70%	64%	74%	47%	4%	0%	76%	63%
11	22%	55%	83%	56%	78%	63%	65%	69%	79%	69%
12	59%	68%	58%	38%	13%	67%	66%	48%	30%	76%
13	75%	60%	0%	68%	41%	46%	65%	65%	63%	77%
14	63%	37%	67%	68%	68%	21%	64%	23%	59%	64%
15	67%	74%	67%	64%	24%	60%	51%	1%	67%	19%
16	8%	75%	69%	1%	54%	70%	56%	63%	61%	59%
17	20%	64%	17%	66%	88%	67%	57%	2%	49%	62%
18	66%	75%	65%	66%	65%	90%	67%	64%	66%	72%
19	64%	66%	74%	62%	71%	69%	53%	65%	68%	13%
20	59%	64%	82%	66%	68%	66%	65%	77%	87%	76%
21	65%	55%	48%	68%	3%	76%	77%	2%	9%	62%
22	65%	33%	67%	65%	1%	64%	63%	81%	53%	53%
23	62%	28%	64%	65%	77%	63%	33%	67%	66%	67%
24	75%	2%	4%	62%	20%	61%	-	-	-	-

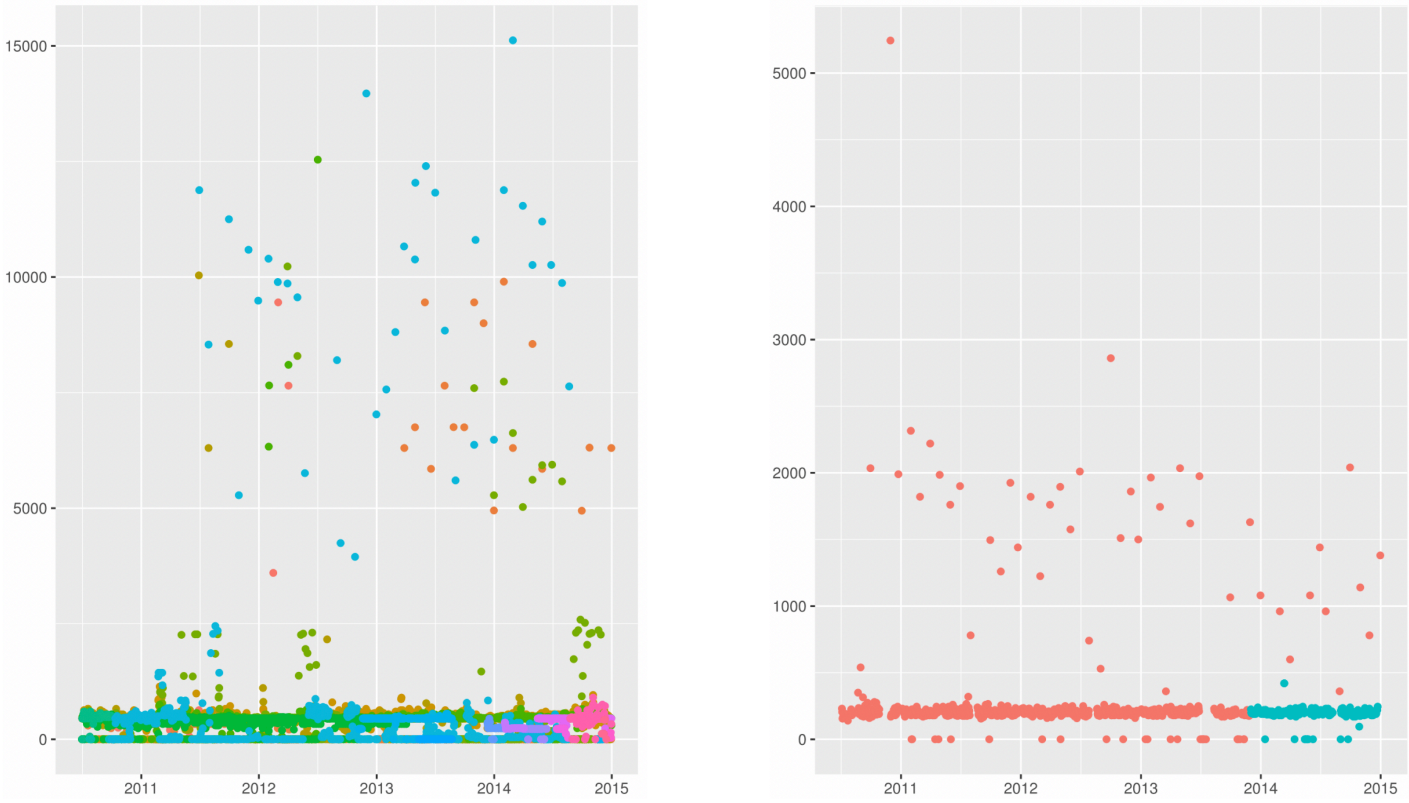
Summary visualization of the gaps in daily log data, by mission. This heat map shows the mean % days into the daily log for each mission, relative to the total possible number of days that could be entered (2010-2014 data).

Start	Monthly Number of Cases	End	Low	High	Mean	Std Dev	Blanks	Zeros	Trend
19502		17265	15150	25072	19903	2612	0.0	0.0	379.2
46		19	3	46	19	9	0.0	0.0	-1.6
156		240	101	326	194	60	0.0	0.0	9.7
16		11	2	76	15	15	0.0	0.0	-2.9
3		13	0	105	9	15	0.0	0.4	-1.8
42		50	25	91	61	16	0.0	0.0	1.2
48		53	34	169	67	25	0.0	0.0	0.6
0		N.A.	0	0	0	0	2.2	9.8	0.0
56		104	34	150	73	25	0.0	0.0	4.6
0		N.A.	0	1	0	0	9.6	1.6	-0.2
195		189	112	544	244	91	0.0	0.0	12.4
0		N.A.	0	0	0	0	6.2	5.8	0.0
127		78	41	132	86	25	0.0	0.0	1.0
0		0	0	4	0	1	0.0	10.4	0.0
0		0	0	0	0	0	0.0	12.0	0.0
0		N.A.	0	0	0	0	2.7	9.3	0.0
0		0	0	0	0	0	0.0	12.0	0.0
0		0	0	4	0	1	0.0	11.3	-0.1
0		18	0	43	9	13	0.0	4.9	7.6
0		0	0	4	0	1	0.0	9.3	0.0
299		236	204	482	323	61	0.0	0.0	-5.0
33		50	15	53	33	9	0.0	0.0	1.1
0		0	0	0	0	0	0.0	12.0	0.0
0		0	0	0	0	0	0.0	12.0	0.0
2		23	1	48	18	11	0.0	0.0	6.3
11		9	1	33	11	6	0.0	0.0	1.2
0		0	0	0	0	0	0.0	12.0	0.0

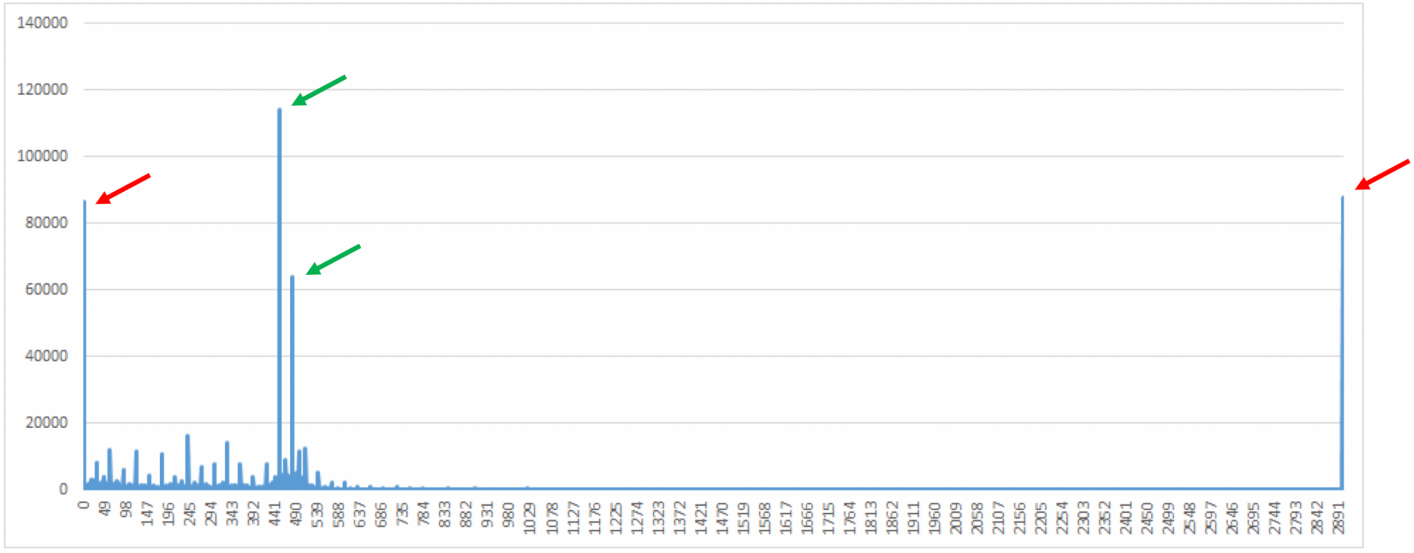
A sample of the summaries of the monthly log data for each mission. The “Blanks” field provides information about the number of months that have no data for that mission, across the years of the dataset reviewed (2010-2014). Top row represents the Grand Total for the entire Consular Network. Subsequent rows represent specific missions.

Entire Dataset Review

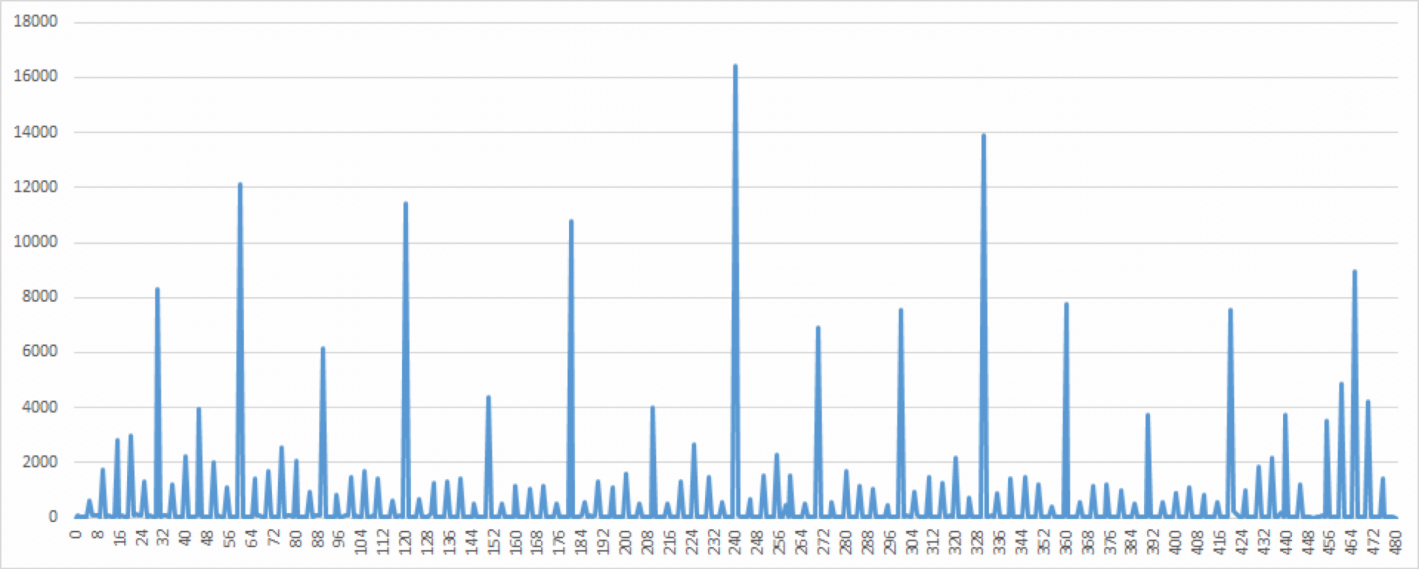
Total daily hours per mission employee (indicated by coloured dots) for 2 missions.



Bar chart of the frequency of specific daily work time values being reported by all employees (entries greater than 2900 mins are put in the same bin). Unsurprisingly, we see peaks around the 7.5-8 hours range (450-480 minutes), but there are also unexpected features: the number of 0 values being reported, and the number of values above 1440 mins (24 hours).

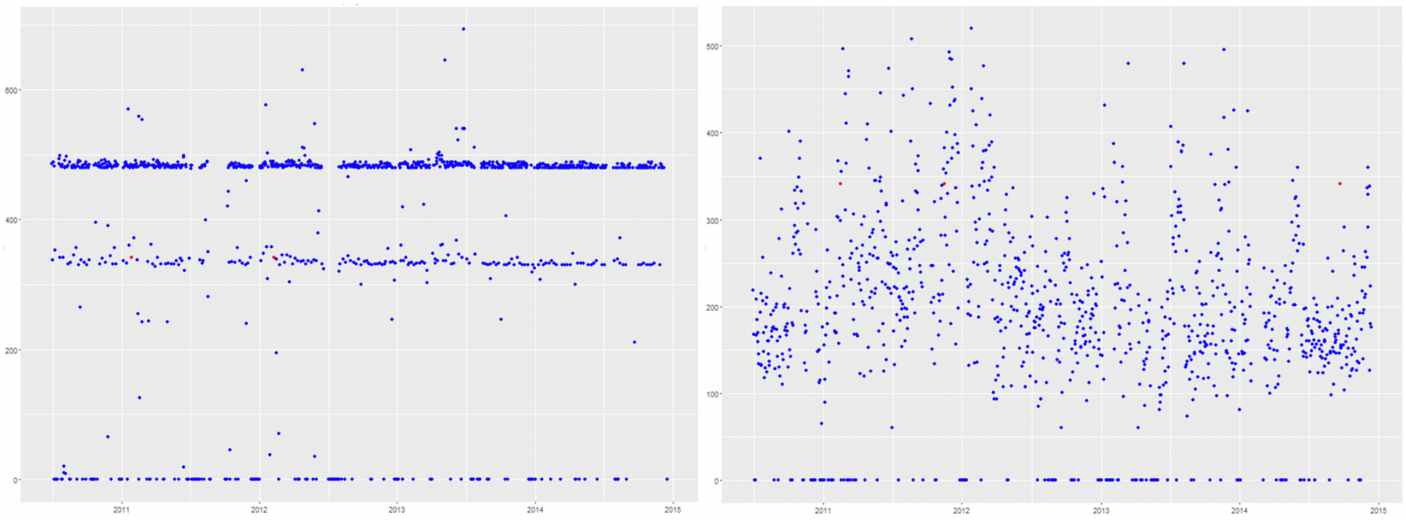


Entire Dataset Review



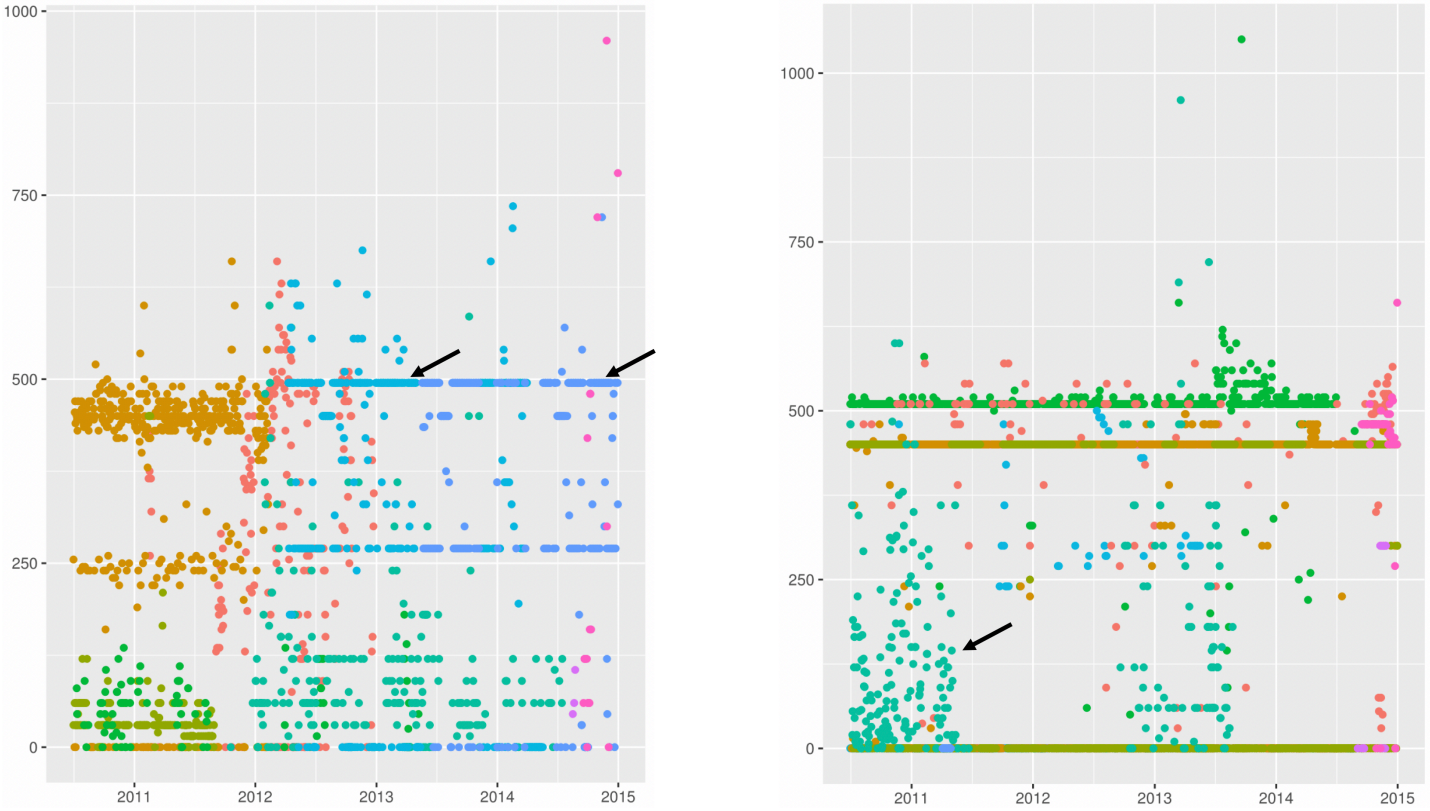
Heaping of work time entries, with counts removed
at 0, 450, and 480+ mins.

Reported daily hours for 2 employees (who reported 371 hours twice, in red);
constant employee on the left; irregular employee on the right.



Entire Dataset Review

Total daily hours per mission employee (indicated by coloured dots) for 2 missions.



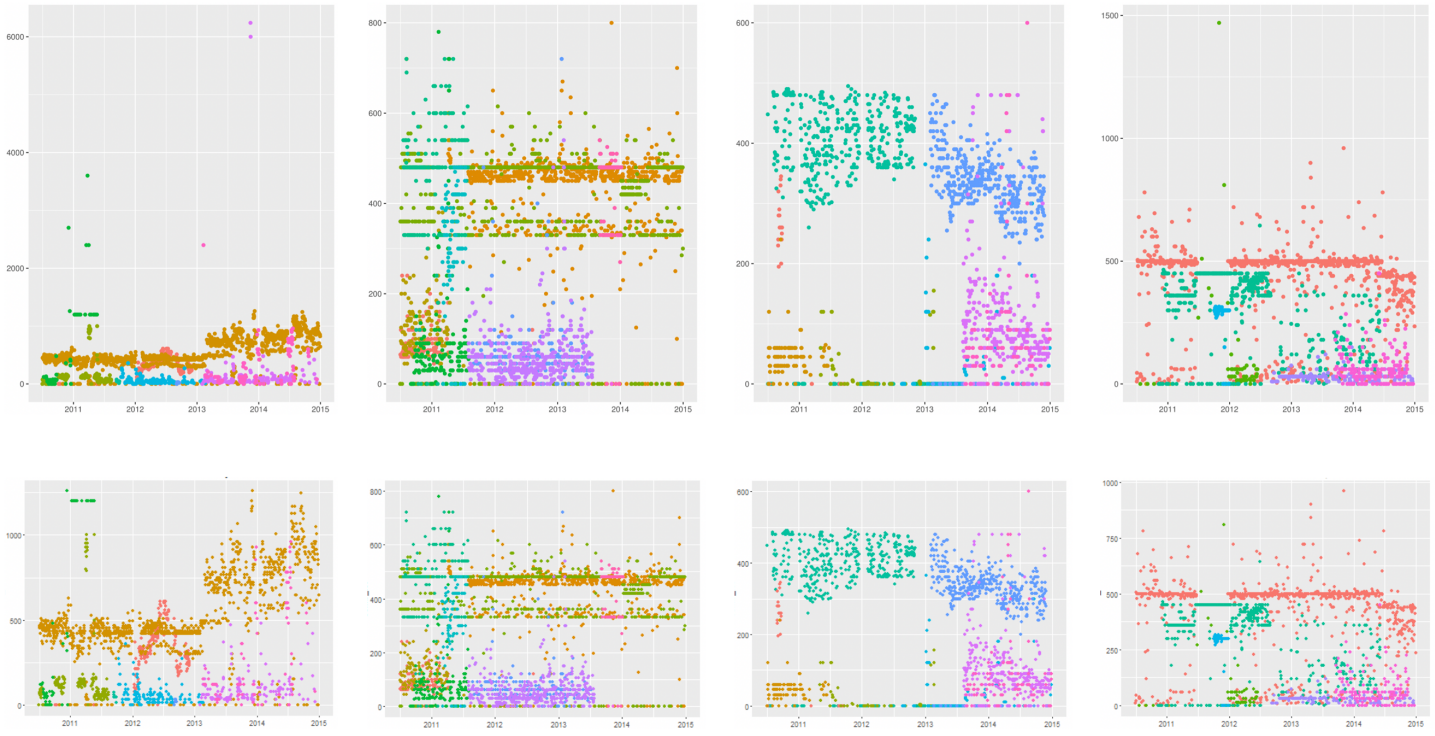
Mission-Level Dataset Review

Small multiples; self-reported daily work times per employee; mission by mission (selection I).

Mission-Level Dataset Review

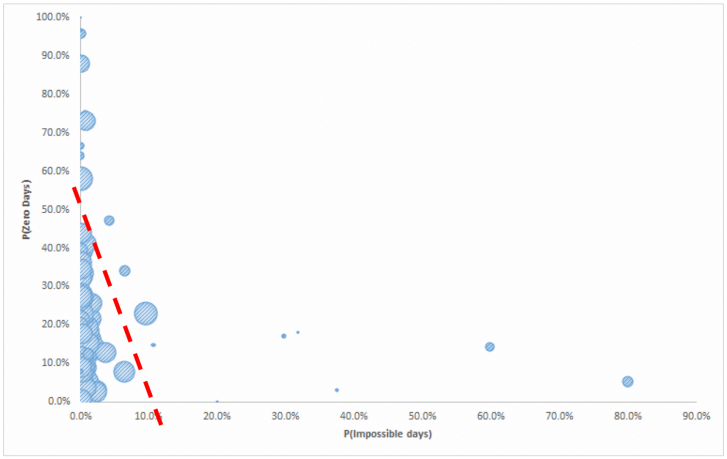
Small multiples; self-reported daily work times per employee; mission by mission (selection II).

4 missions with anomalies (top); without (bottom).

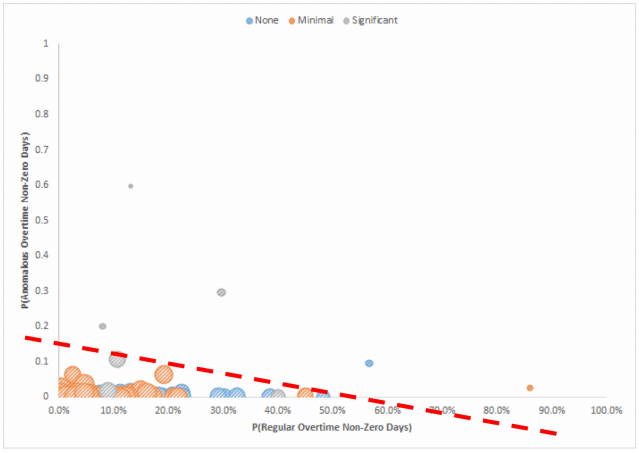


Plausibility of Work Hours

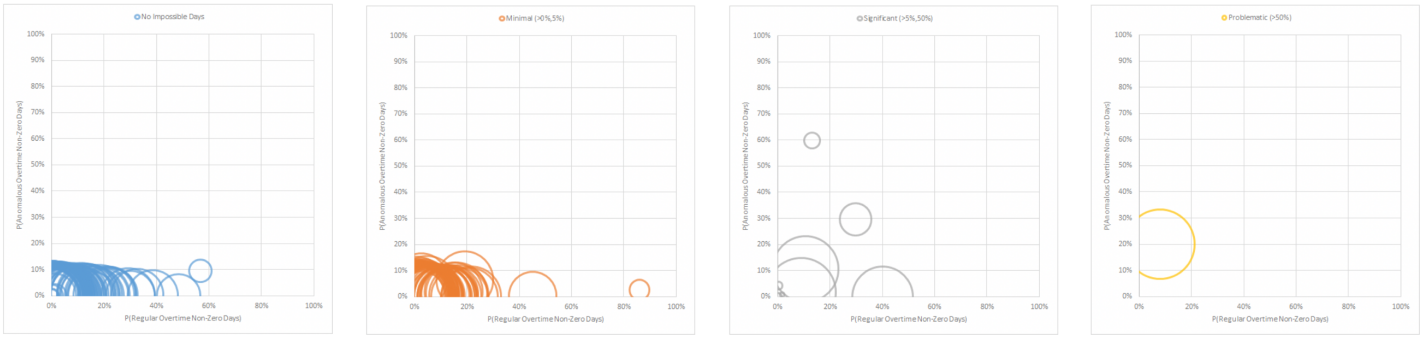
Proportion of Zero Days against Proportion of Impossible Days, per mission (size related to number of entries per mission)



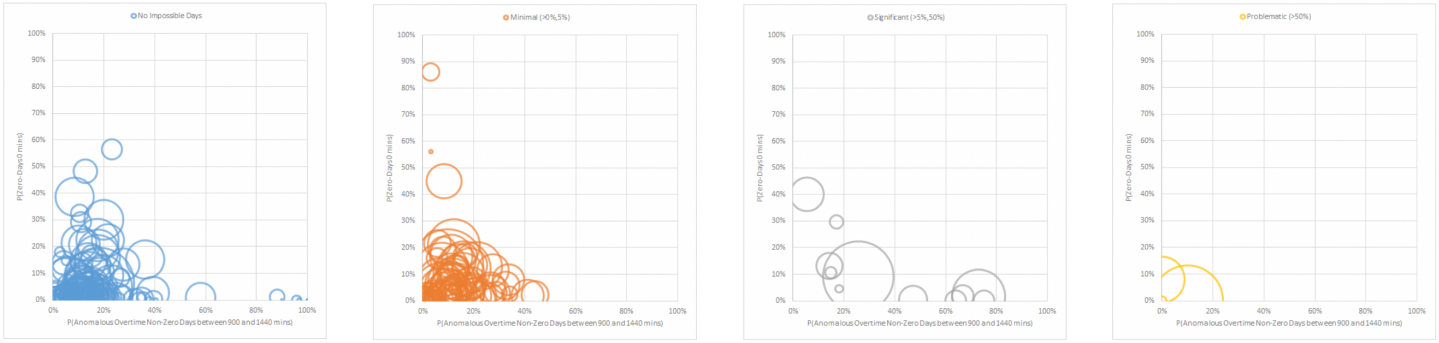
Proportion of Anomalous Overtime Days against Proportion of Plausible Overtime Days, per mission (size is related to number of entries per mission; colour to Proportion of Impossible Days)



Proportion of Anomalous Overtime Days against Proportion of Plausible Days, by mission and Proportion of Impossible Days (bubble size is linked to number of entries per mission).

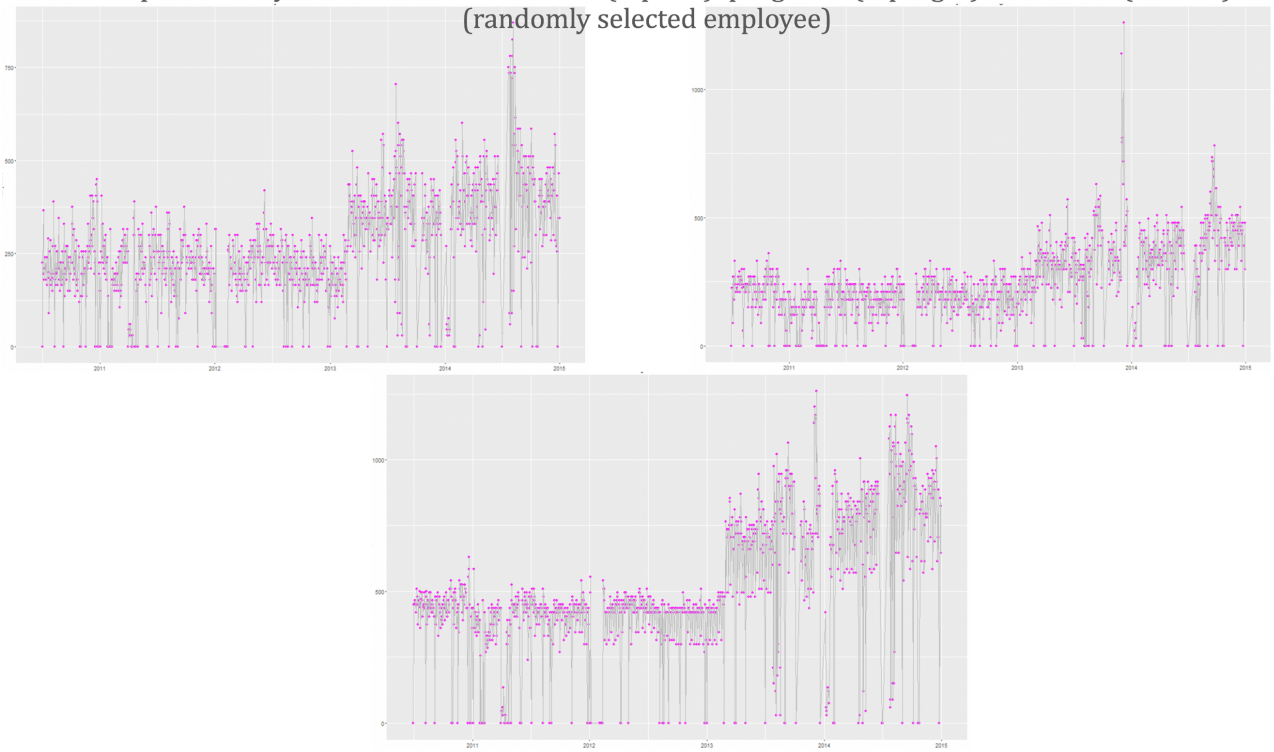


Proportion of Zero Days against Proportion of Anomalous Overtime Days, by mission and Proportion of Impossible Days (bubble size is linked to number of entries per mission).



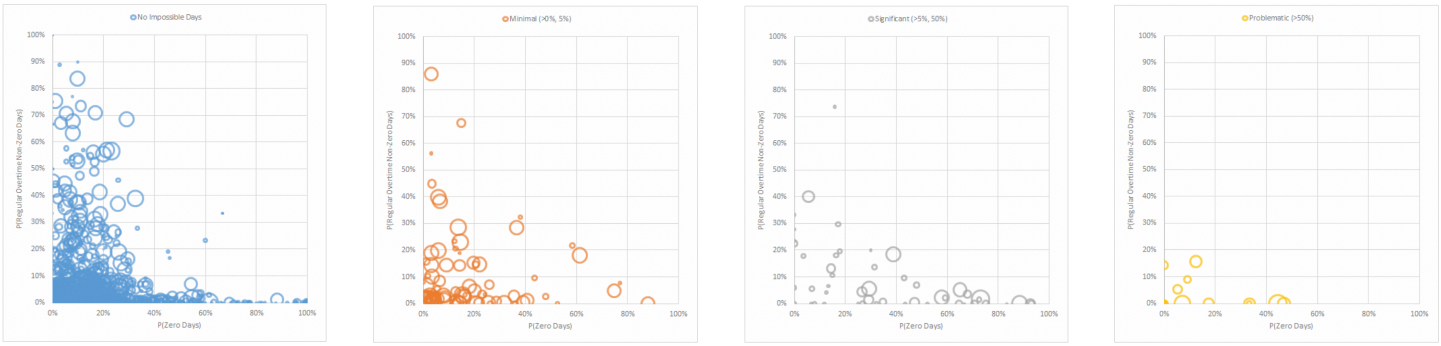
Employee-Level Dataset Review

Self-reported daily times; cases and services (top left); programs (top right); combined (bottom) (randomly selected employee)



Plausibility of Work Hours

Proportion of Anomalous Overtime Days against Proportion of Plausible Days, by employee and Proportion of Impossible Days (bubble size is linked to number of entries per employee).



Proportion of Zero Days against Proportion of Anomalous Overtime Days, by employee and Proportion of Impossible Days (bubble size is linked to number of entries per employee).