Canadian Consular Network

Practical Data Visualisation

P. Boily

Contents

Project Description

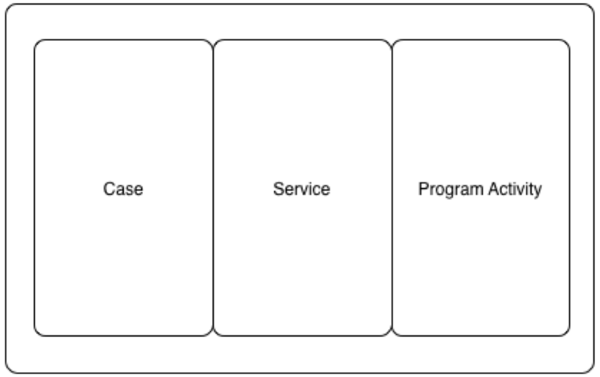
Project Description
Context

- Within *Global Affairs Canada* (GAC), *Consular Corporate Management and Innovation* (CCMI) has a software application (COSMOS) 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.
- COSMOS 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 stretches back over approximately twenty years.
- It is currently used to
  - determine the **effectiveness** of mission consular programs,
  - identify **weaknesses** to be resolved through HR, training and other solutions, and
  - 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). GAC would like to determine if the **current system meets their needs**.

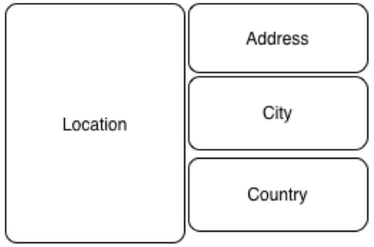
Project Description
Consular System

Personal Details
Type of assistance required
Wait time
Assistance Time
Type of assistance provided

Personal Details
Place in Organizational Structure
Role
Group Membership
Employee ID
Time Worked (hours, days)
Cases Worked On
Services Worked On
People worked with
Clients worked with



Activity ID
Activity Type or Category
Date and Time Activity Opened
Date and Time Activity Closed
Duration of Activity
Times when Activity was actively worked on
Employee who opened Activity
Employees who worked on Activity
Employee who closed Activity
Client owner of Activity
Activity Difficultly Level



City Properties
Country Properties
(e.g. Geography, Population)
Government Status
Many other properties, metrics and
measures (e.g. Corruption Index,
Canadian Traveller Statistics)

Mission Name
Mission ID
Mission Type
Contact Notes
Location
People
Functions
Mandate/Mission Statement
Work
Organizational Structure
Hours open
Years open
Financial Aspects of Mission
Phonebook
Calendar
Other Metrics (e.g. Hardship Rating)

Project Description
Data

- Data of **primary interest** for consular management is contained in 4 COMIP tables
  - logs of mission activities (**cases**, **services**, and **programs**)
  - **time spent** on these mission activities, **daily** and **monthly**.
- Within these tables, data is available across a time span of **10 years**, from 2005 – 2014.
- During a system upgrade in 2010, the categories relating to cases and services were changed, resulting in a **break in the dataset** at this time.
- Discussions with CCMI suggests that **data accuracy** is highest from mid-2010 onwards.
- The focus of this analysis is on the **reliability** of case, service, and program data collected between July 2010 and December 2014.

Data Reliability

Data Reliability

- In data analytical endeavours, the **quality of the output** is affected by the **quality of the input**, especially when it is **self-reported** (which is the case with COMIP).
- CCMI understands (domain expertise!) 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**
  - not a requirement in order to produce a monthly log.
- **BUT** abandoning daily log data entered into the system is a problem as it is impossible to create monthly log data that **accurately reflects the reality** of monthly work in the mission without (some) information from employees about their daily work during the month.
- Daily data is being used *de facto* to create the monthly logs.
- Thus, while certain types of consular data analysis may be conducted using monthly aggregates, **data validation** has to occur at the daily data level.

Data Reliability
Basic Checks

- Basic data checks were carried out using standard **mathematical** and **logical** tools to verify consistency and completeness of the dataset.

- General findings:
  - data is **clean** (e.g. fields had valid values in expected ranges);
  - there were significant **gaps in data entry**;
  - as well as **logical inconsistencies** resulting from data entries issues.

- **No incorrect data types or ranges:** the data entered into fields was consistent with the designated data types and ranges for each field
  - non-negative numeric value for fields representing time spent on cases and services, or number of such cases and services
  - character values for fields providing categorical information, etc.

Data Reliability
Basic Checks

- **Missing values:** some case and service category related fields were **empty** as the data entry system allows employees to leave fields blank, rather than fill in a "0" or other numbers. Given this context, an empty cell could be interpreted as implicitly indicating a "0".
  - **challenge:** inability to distinguish between a field that could have had a value, **in principle**, and a field that was necessarily "0", because the mission in question **does not provide that type of service**, say.
  - does not allow for the possibility that work was carried out and never entered.
  - **solution:** empty cells given the value "NA" rather than "0".
- **Change to Fields:** between 2010 and 2011 a change was made to the activity categories used to collect data about case and service activities. This change in categories was not a **straightforward** splitting or combining of previous categories.
  - old categories were not mapped onto new categories
  - for time fields, combining all activity times into a single time estimate across for **long-term** activities

Data Reliability
Basic Checks – Data Gaps

- A **gap** in the data is a date for which there is no corresponding row entry.
  - for a given mission & month, if there are 31 days in the month and only 20 days with data in the mission's daily log file, there are **11 data gaps**.
- The system does not enforce a data entry for every day, hence gaps in the data may represent:
  - a day where **work was done but not logged**, or
  - a day where **no work was done** (mission closed, or mission open but no cases were opened or services were rendered)

	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).

Data Reliability
Basic Checks – Data Gaps

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.

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

Data Reliability
Logical Inconsistencies

- There were a significant number of inconsistencies in the **daily vs. monthly** logs
  - some monthly logs had no associated daily logs
  - some monthly logs had associated daily logs that did not "add up" (above or below, for one or multiple fields)
- There were other discrepancies relating to **case/service numbers** and **time worked**. Some of the values in these tables indicate that the data entered within the logs is not consistent:
  - for example, a daily log entry showing that cases have been **opened on that day**, but also showing that no time has been spent on those activities.

Citizenship Related		Time = 0	Time > 0	Total
Case = 0	Service = 0	5066	974	6040
	Service > 0	39	7916	7955
Case > 0	Service = 0	13	50	63
	Service > 0		1242	1242
Total		**5118**	**10182**	**15300**

Data Reliability
Entire Dataset Review

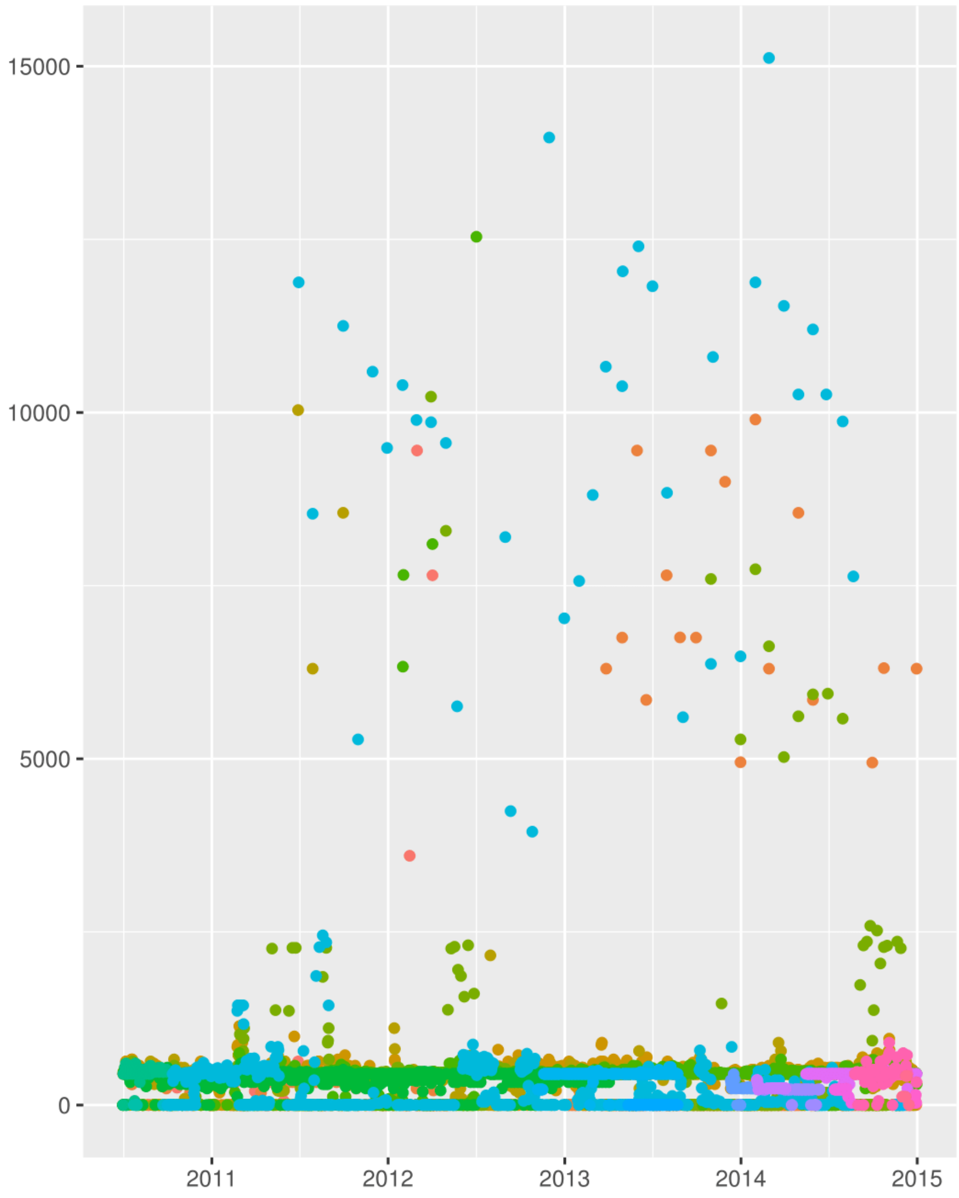
- Focus of this full dataset review was on **employee-hours worked**, which revealed a mixture of data patterns which were either **improbable** or **difficult to explain**.
- Time series plots can be used to detect some anomalies in data, such as recorded **daily working times greater than 1440 minutes** (24 hours)
  - may be due to multiple employees' hours entered under a single employee ID
  - could also be caused by data entry issues
  - consider the two missions (next page) which are very **noisy** due to a large number of the time entries being greater than 1440 minutes (notice the "horizontal lines")

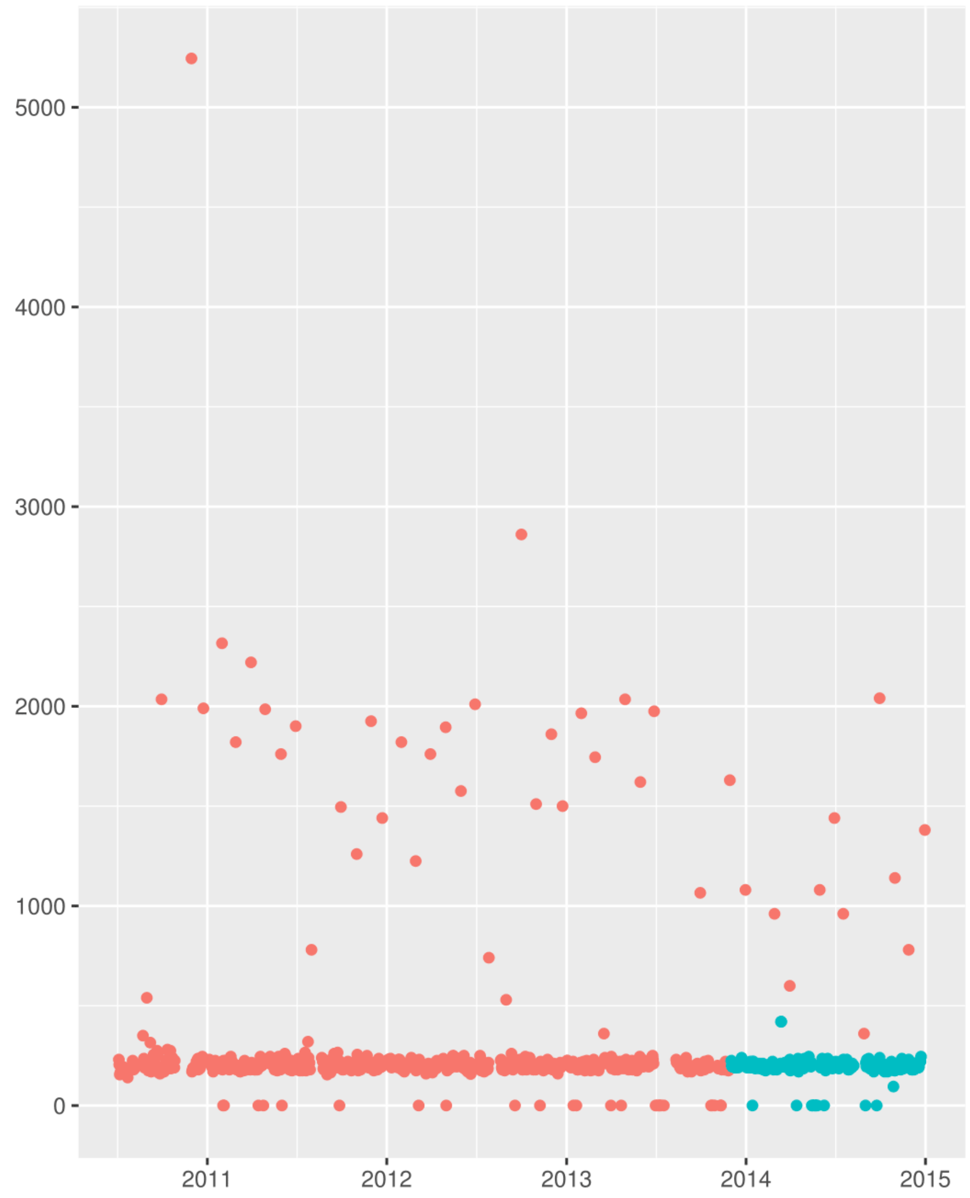
Proportion of impossible days, per mission and per employee.

The vast majority of missions and employees never enter an impossible number of minutes.

Impossible Days Type	Missions	Employees
None (0%)	143	1538
Minimal (>0% to 5%)	79	82
Significant (>5% to 50%)	11	49
Problematic (>50%)	3	15
Total	**236**	**1684**

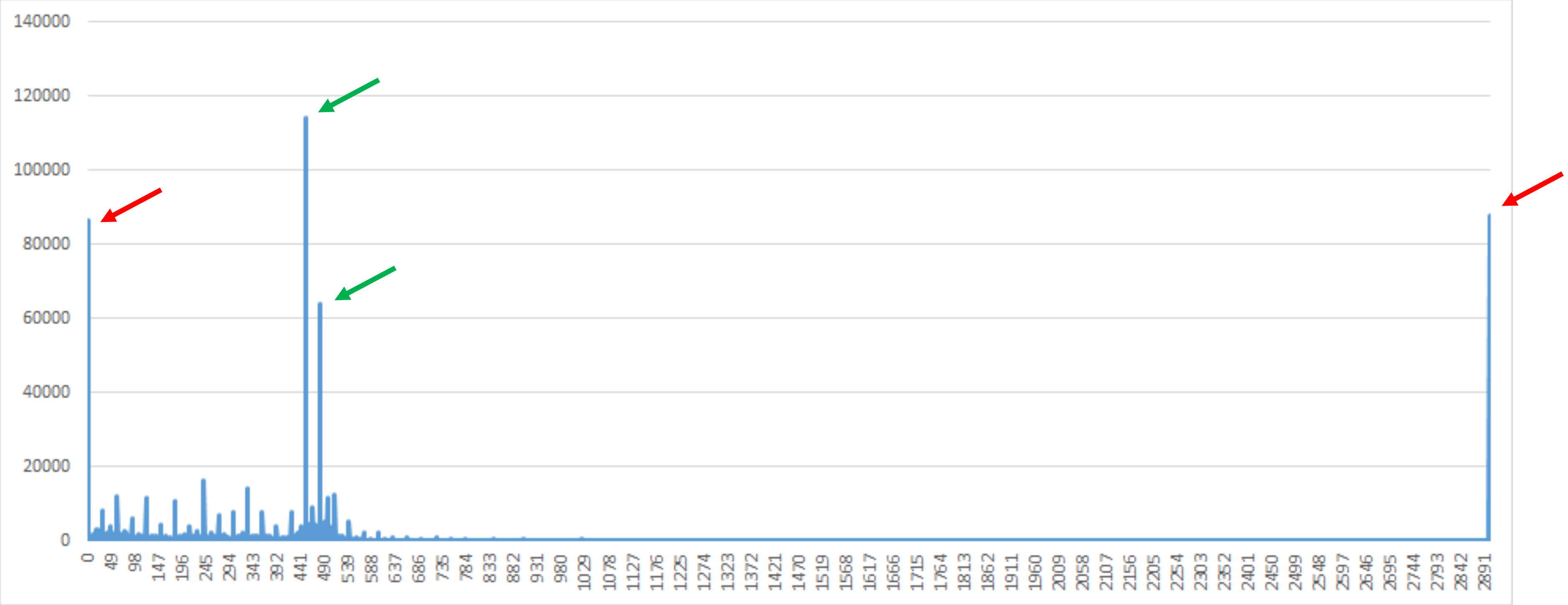
Total daily hours per mission employee (indicated by coloured dots) for 2 missions.





Data Reliability
Entire Dataset Review – Reported Daily Work Times

- Another basic check is to plot a 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).



Dataset Reliabilty

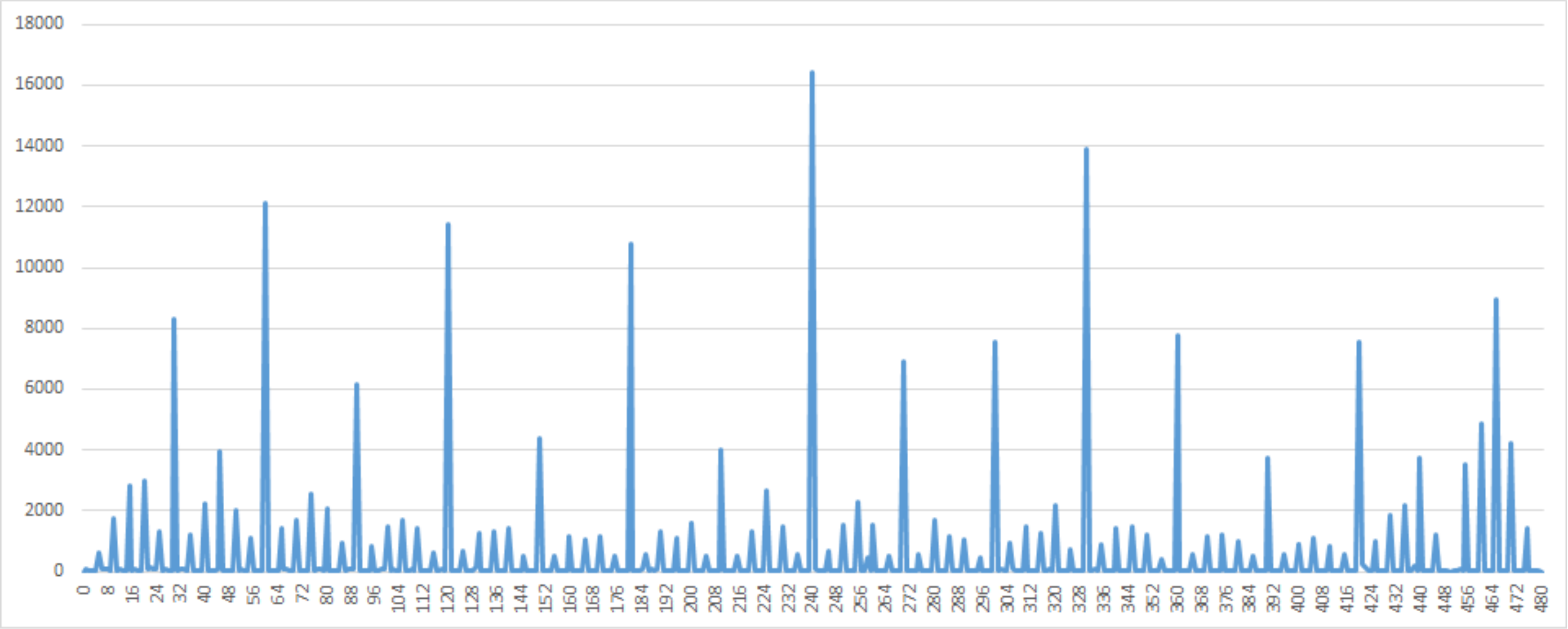
Entire Dataset Review – Reported Daily Work Times

- Basic statistics **corroborate** the overall picture:
  - MEDIAN in the right ballpark, as are 1st and 3rd QUARTILES
  - large number of "0" entries bring the mean down (in spite of large number of impossible entries)
  - MAX is simply ridiculous
  - STD DEV is really high (again, probably driven by large number of "0" and impossible entries)

Statistic	**Value (mins)**
MIN	0
1st QUARTILE	100
MEDIAN	390
3rd QUARTILE	465
MAX	42710
MEAN	321.89
STD DEV	389.44
N (ENTRIES)	590,085
N (MISSIONS)	236
N (EMPLOYEES)	1684

Data Reliability
Entire Dataset Review – Heaping

- Another issue that can be detected fairly easily is the **heaping** of worktime entries: psychologically, human beings are more likely to report by rounding to the nearest 60-, 30-, 20-, 15-, 10- or 5- minute blocks.

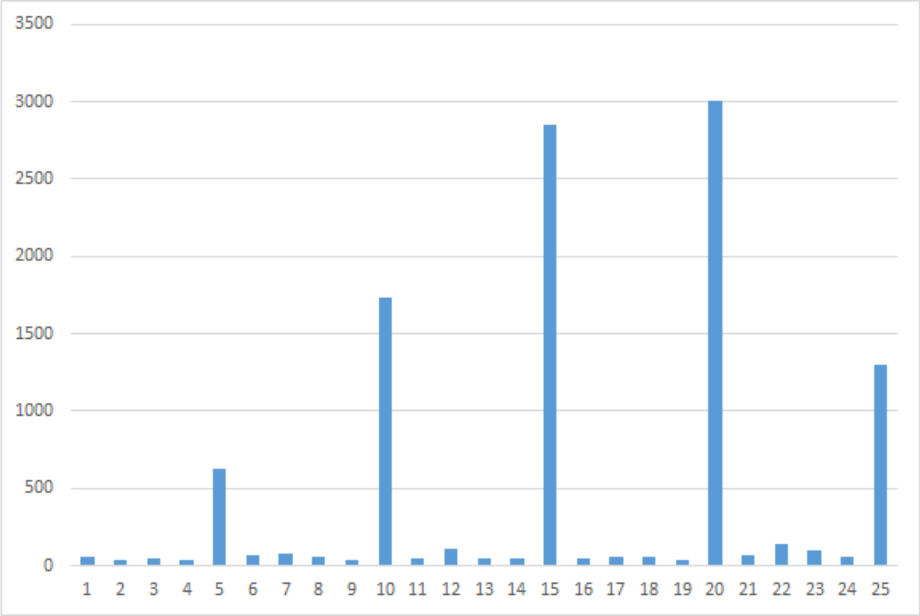


With counts removed at 0, 450, and 480+ mins.

Data Reliability

Entire Dataset Review – Heaping

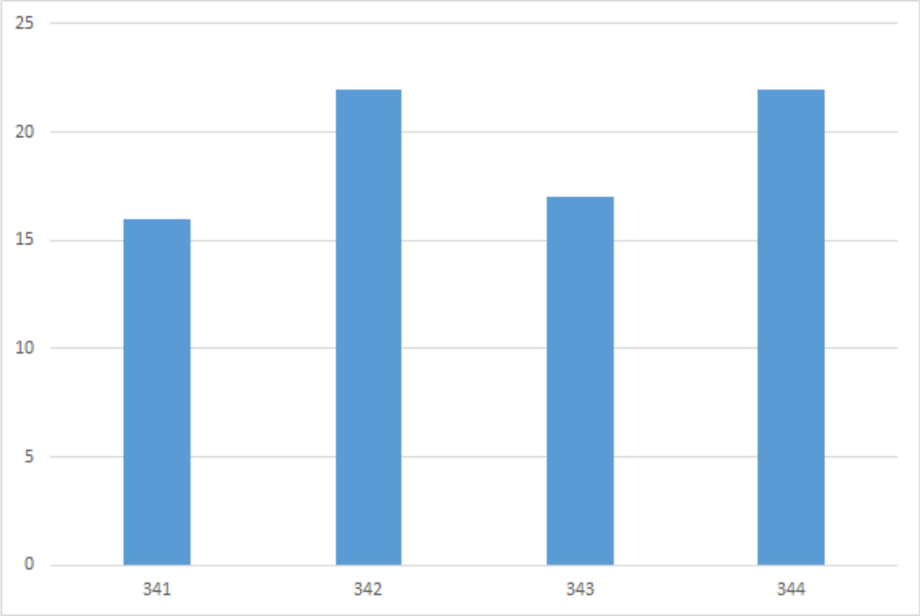
- Anticipated heaping is accompanied by **oddly specific reported times**, which should cause analysts to question the validity of some of the entries in the 25-minutes-and-under range:
  - Did any employee really work **1 minute** on an activity on select days? Is this a typo? Was there a misunderstanding of the units – did the employee think that 1 stood for 1 hour, or 1 day? While these are somewhat infrequent (at least compared to the total number of entries), they do raise doubt as to the **validity** and **reliability** of the reported numbers for the dataset as a whole.
  - For instance, do 7.5 hrs and 8 hrs appear repeatedly **because these are the expected number of hours to be recorded**, or the actual number of worked hours?



Over the first 25 minutes.

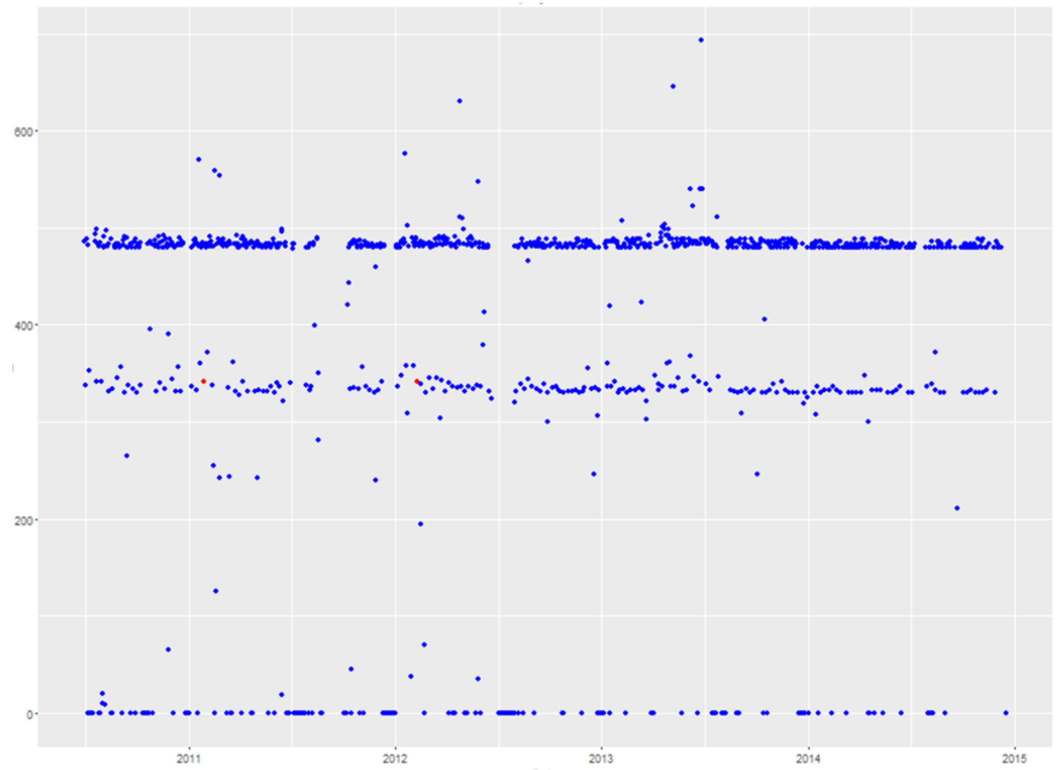
Data Reliability
Entire Dataset Review – Anomalies

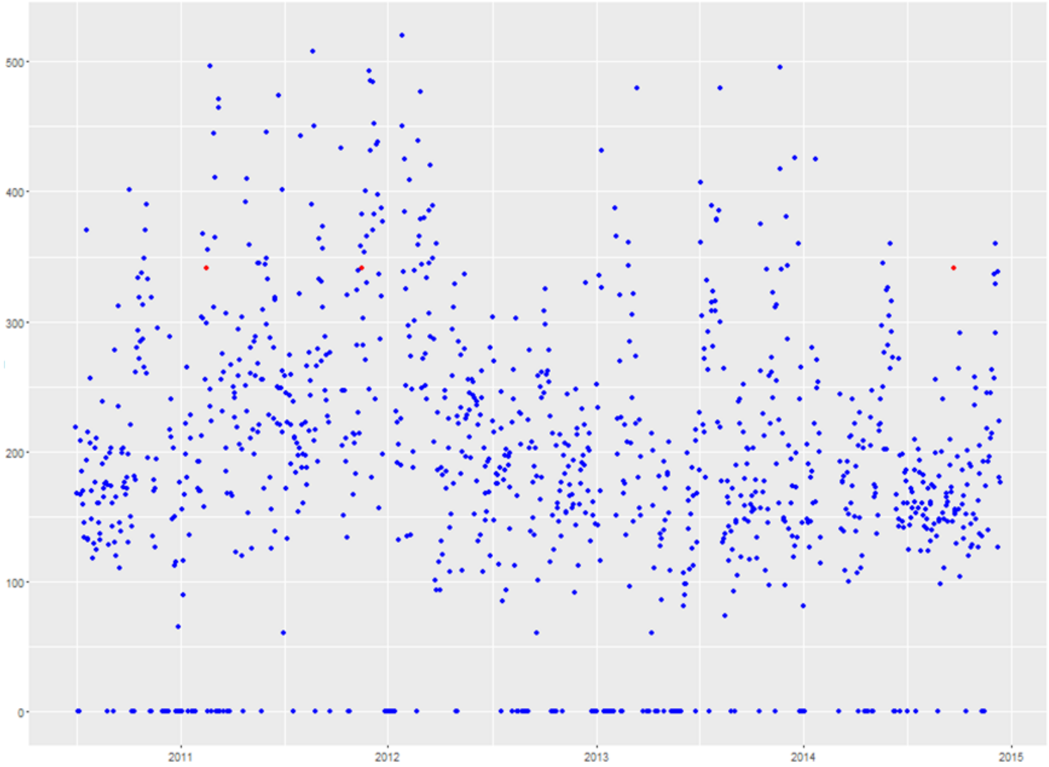
- Consider the fragment of the chart running from 341 to 344 mins.
- It is conceivable that an employee who reported working 341 minutes on a given day **did indeed work** 341 minutes, but it is also possible that such reporting masks **backtracking** from monthly estimates to daily reports.
- On the next page, we see two different employees from two different missions who each reported working 341 minutes twice. The time series for each employee are **vastly different** but they do not allow us to differentiate between the two alternatives.
- Without **external audit data**, we cannot differentiate between valid and potentially invalid data.



From 341 to 344 minutes.

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

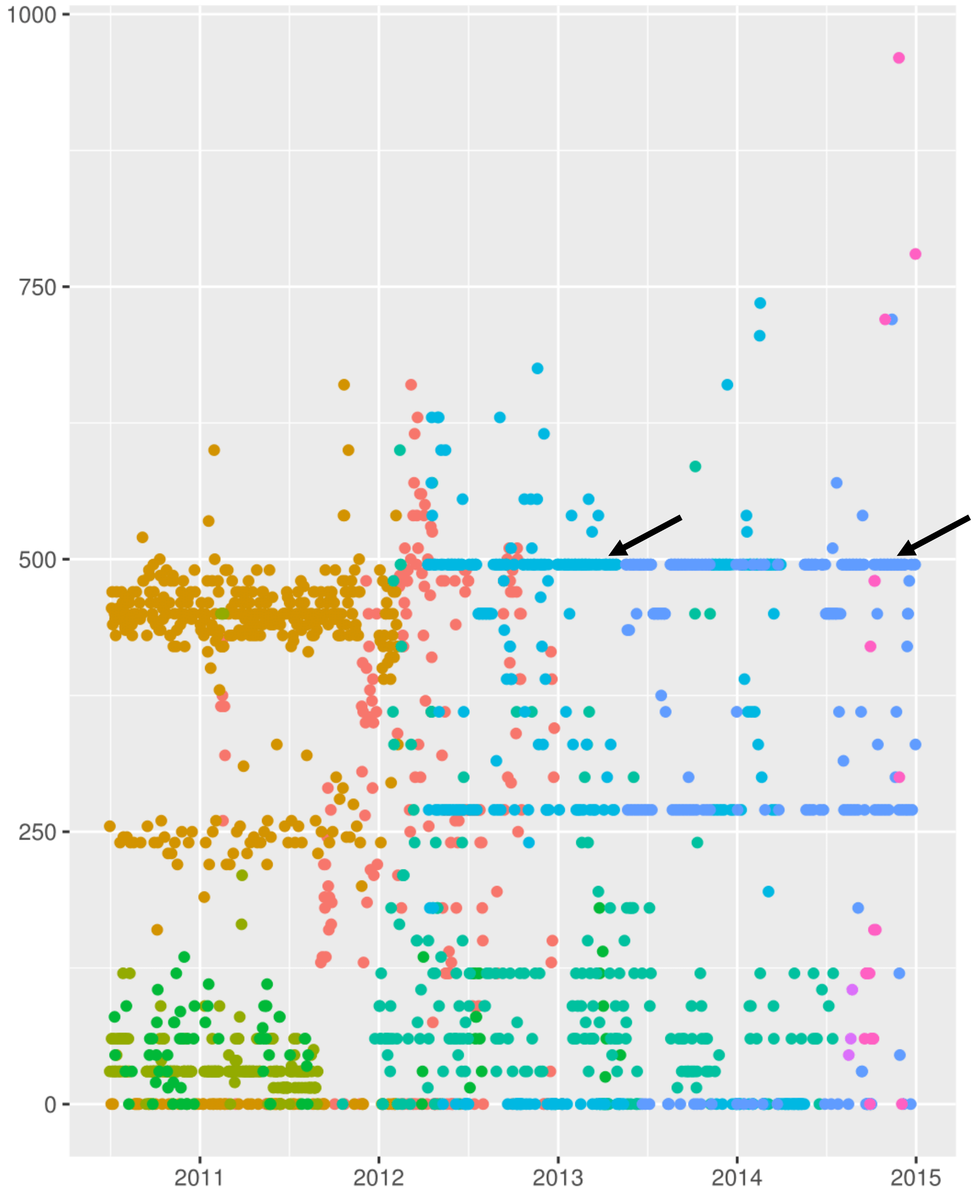


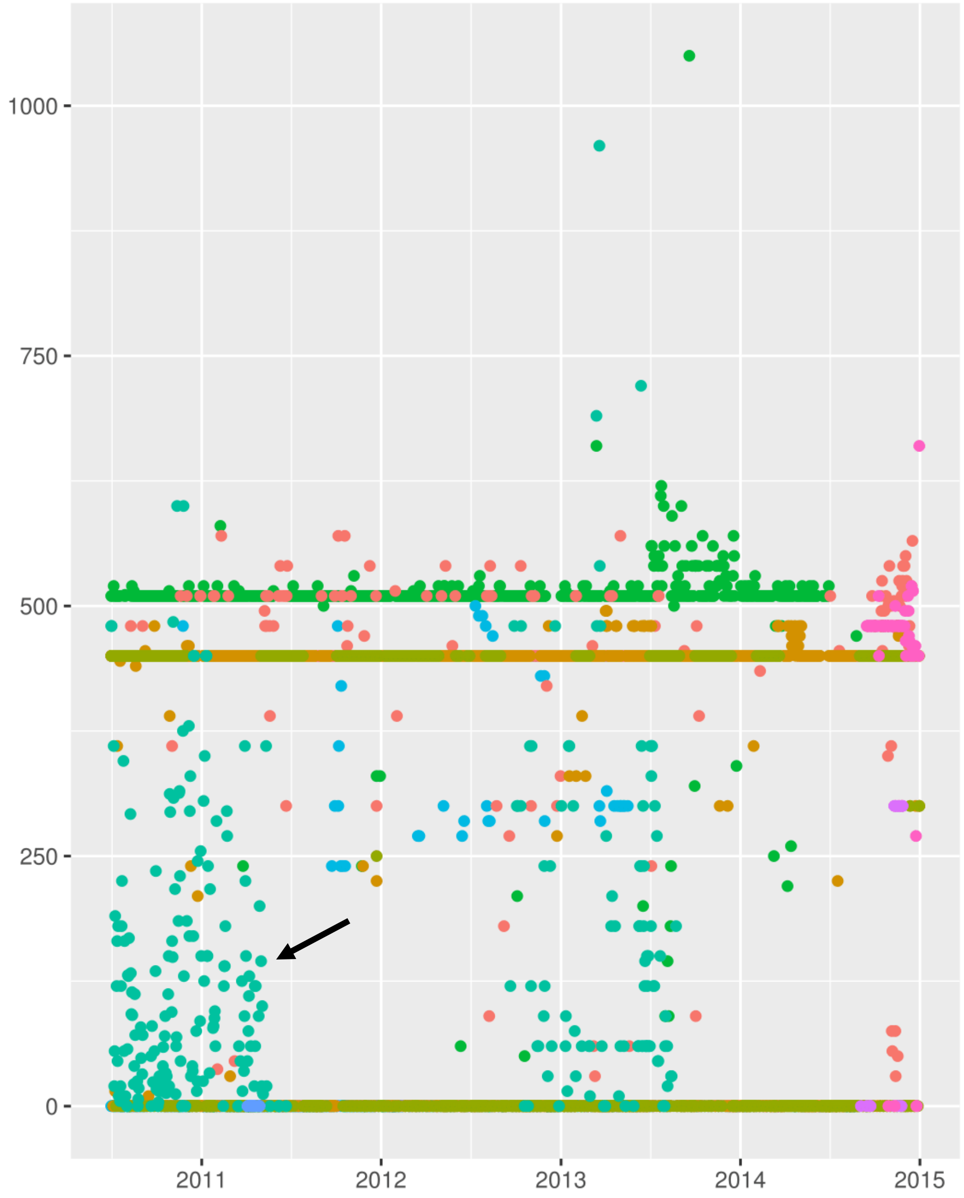


Data Reliability
Entire Dataset Review – Anomalies

- It is difficult to identify patterns which are **universal** to all missions.
  - For instance, there are two employees of one mission (identified by the arrows, next page, left) who seem to work from **Sunday to Thursday** more or less every week, with roughly the **same time range** (between 270 minutes and 480 minutes), with a **shorter day** on Thursdays.
  - In another mission, all employees seem to record **roughly the same amount of time** on every working day (but which may differ from one employee to the next), except for one employee who seems to play the role of a **gopher** (next page, right).
- Ideally, the **lack of similarity** between these two missions (as well as the missions previously shown) would be an indication that at least **one of those contains invalid data**
  - but not every mission provides the same services or has, *a priori*, similar traffic patterns
  - this external data is not available in the database so there is no basic check to determine the validity of patterns
- In fraud detection, non-compliance with **Benford's Law** is sometimes used to identify anomalous observations, but the structure of the data and the overall preference for entries starting with 4 – such as 450 and 480 – make it inapplicable to this case.

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





Data Reliability
Entire Dataset Review – Anomalies

- Looking across the entire dataset, there are are situations where the data indicates that something **unexpected**, **improbable**, or **difficult to explain** is occurring.
- By doing a survey of data **across all missions**, we can **detect** and **catalogue** these anomalies.
- Unfortunately, while in some circumstances these types of occurrences might indicate **invalid** data, in the current self-reporting collection system, there are generally other **conflating explanations** which cannot be disentangled from this question of validity.
- Common data analysis techniques can be used to detect a variety of unusual or anomalous patterns, but **interpreting** these findings is made challenging by the way that information is collected by GAC, along with what type of information is collected.
- A detailed discussion of possible scenarios at the mission level will make this all the more apparent.

Data Reliability
Mission-Level Dataset Review

- The question of validity of the data at the **mission level** is simpler to tackle in theory.
- The number of employees at each mission is **small**, and there was some hope that a mission's employees would **all follow roughly similar reporting paths**, even though these could be **different from mission to mission**.
- Given the logical discrepancies found in the data relating to cases, services and time worked, and the fact that cases were only ever identified as open (without **duration** or **closure**), it was decided that the most logical variable with which to work when analysing mission data remained **the combined time spent by each employee on cases, services, and programs, on a daily basis**.
- A **stand-alone validity metric** could not be constructed, because **domain expertise** about each mission (and potentially each of its employees) is required in order to differentiate and identify which of various interpretations is most likely to explain the observed data.

Data Reliability
Mission-Level Dataset Review

- A three-pronged approach was taken in an effort to detect **relevant and useable patterns**.
- First, a detailed, **highly granular** review of data for each mission was carried out by creating time series of the daily mission employee-hours, over the entire range of data (2010-2014).
  - In principle, the detection of consistent employee-hour patterns in this data could allow **baseline behaviour patterns for the mission** to be developed, against which anomalous data entry patterns could be detected.
- Second, a wide variety of **possible data entry scenarios** were generated in order to establish the data entry patterns that would be created by **entry of invalid data**.
  - As importantly, the attempt was then made to **distinguish** these from patterns created by entry of valid data.
- Third, the mission data was assessed by analyzing the **overall plausibility** of the work hours entered for the mission.
  - This was accomplished by dividing the daily time worked by employees into several categories related to plausibility of the amount of time entered (e.g. **full working day**, **overtime**, **more-than-available-hours**) and examining the patterns and relationships between these categories within a mission.

Data Reliability

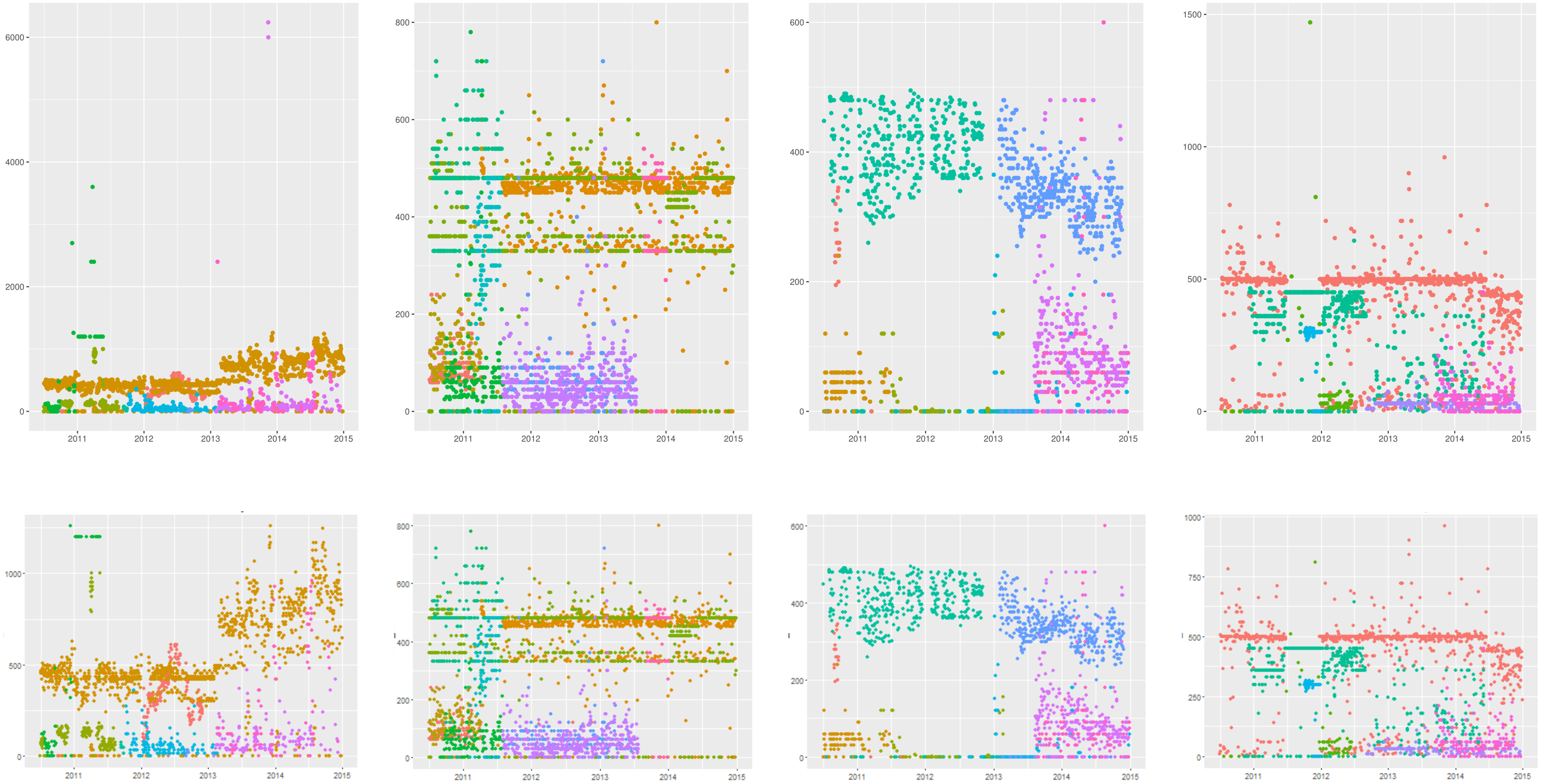
Mission-Level Dataset Review – Baseline Consistency Analysis

- Discussions with CCMI suggested that variability both **within** and **across** missions could realistically be expected to be **high**, which made it likely that searching for anomalous patterns in the data would be very difficult.
- This suspicion was confirmed by an **extensive manual analysis** of the employee work hour time series for each mission.
- A **visual analysis** of these time series was used to detect the potential for a baseline description of mission data (see small multiples on the next pages).

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

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

Missions with anomalies (top); without (bottom).



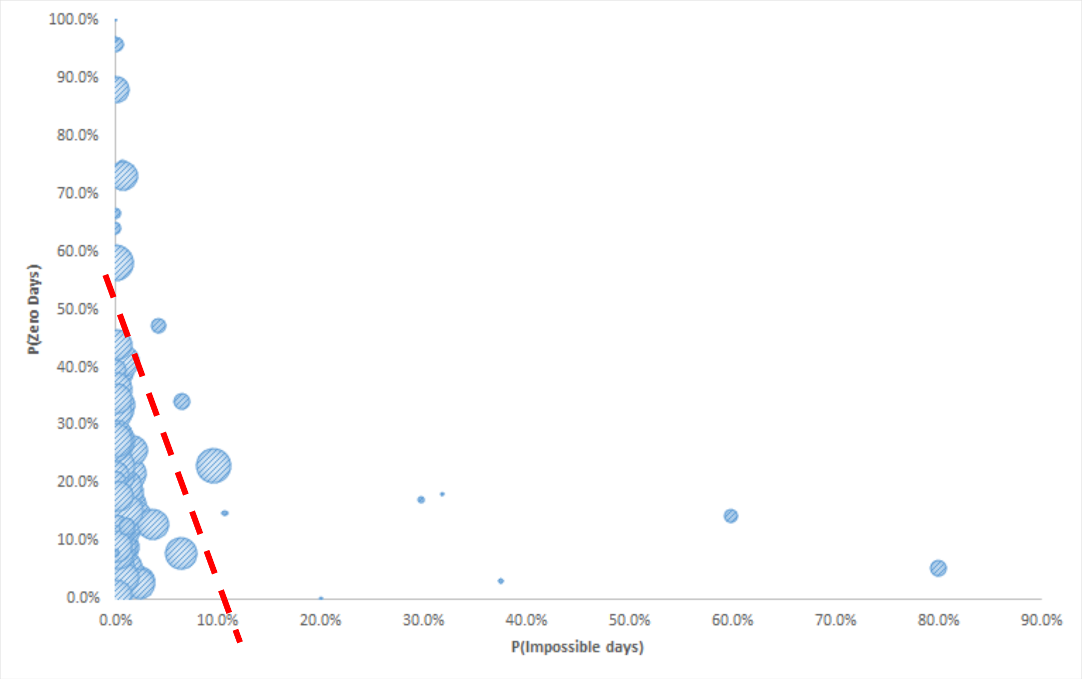
Practical Data Visualisation

CANADIAN CONSULAR NETWORK

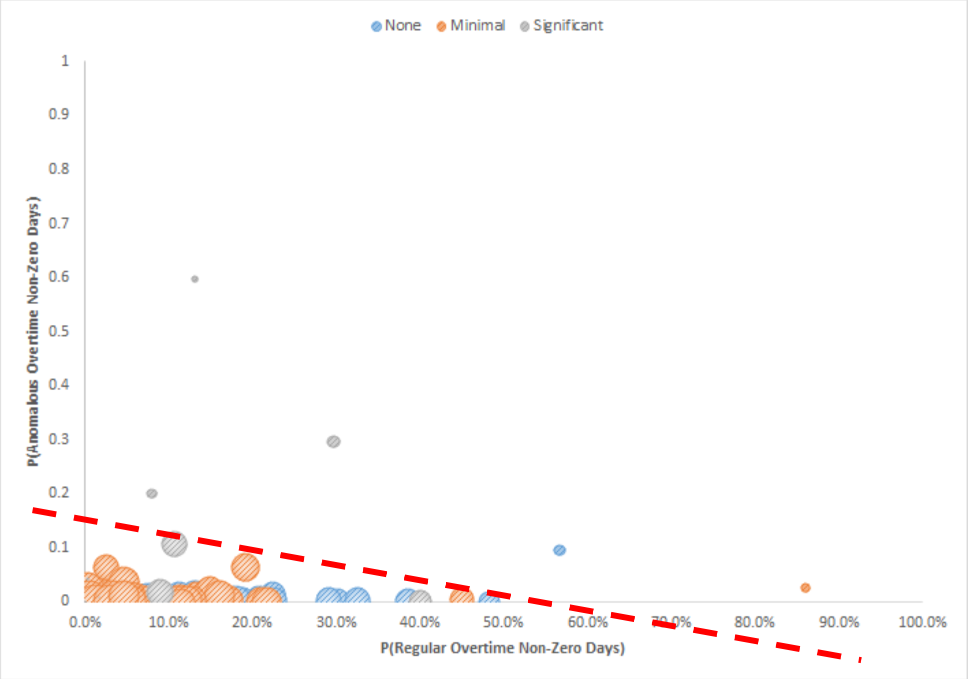
Data Reliability
Plausibility of Work Hours

- To analyse **plausibility** of the daily self-reported work time, we can start by studying the relationship between the number of entries for which:
  - **no time** has been recorded;
  - an **impossible** amount of time has been recorded (1440+ minutes; 24+hrs);
  - a **reasonable** amount of time has been recorded (0-500 minutes; 0-8hrs20min);
  - a **plausible** amount of overtime has been recorded (500-900 minutes; 8h20min-15hrs), and
  - an **anomalous** amount of overtime has been recorded (900-1440 minutes; 15hrs-24hrs).
- **Bubble charts** might help identify anomalous missions.

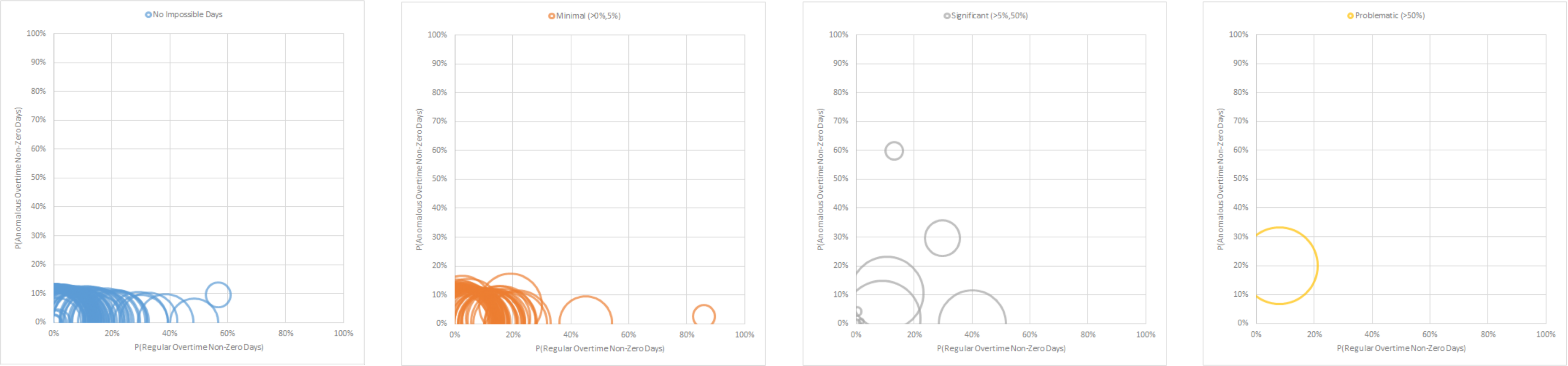
Proportion of Zero Days against Proportion of Impossible Days, per mission (size related to number of entries per mission)



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)

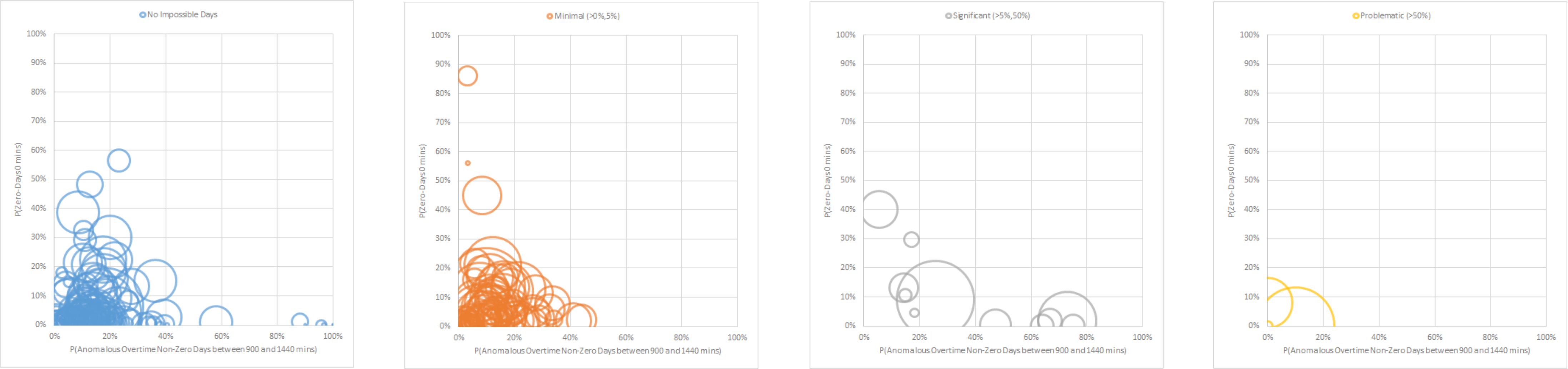


Proportion of Anomalous Overtime Days against Proportion of Plausible Days, by mission and Proportion of Impossible Days (bubble size is linked to number of entries per mission).



Bubbles “rise” as Proportion of Impossible Days increases.

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).

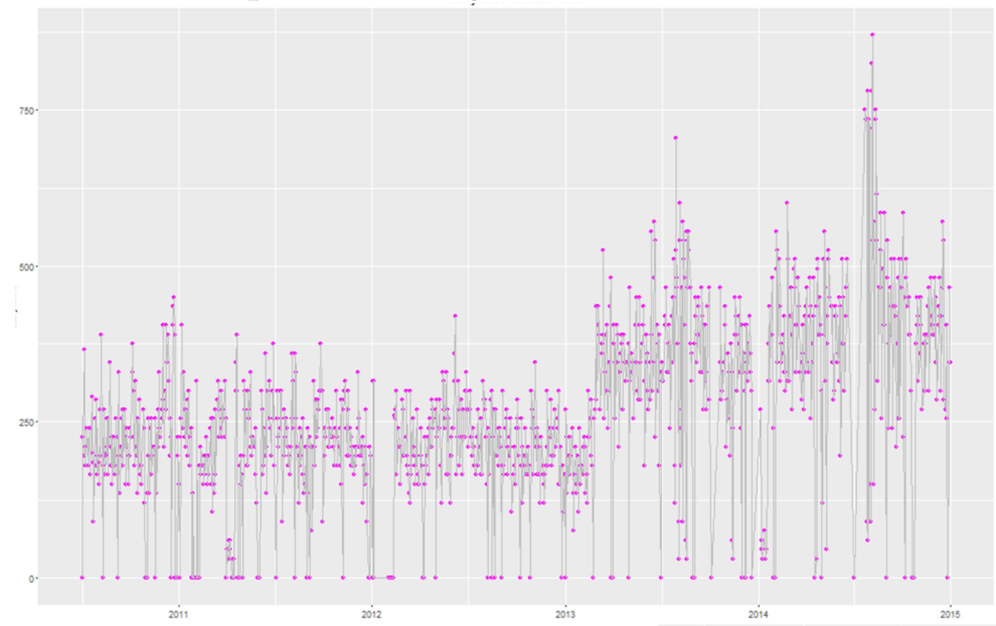


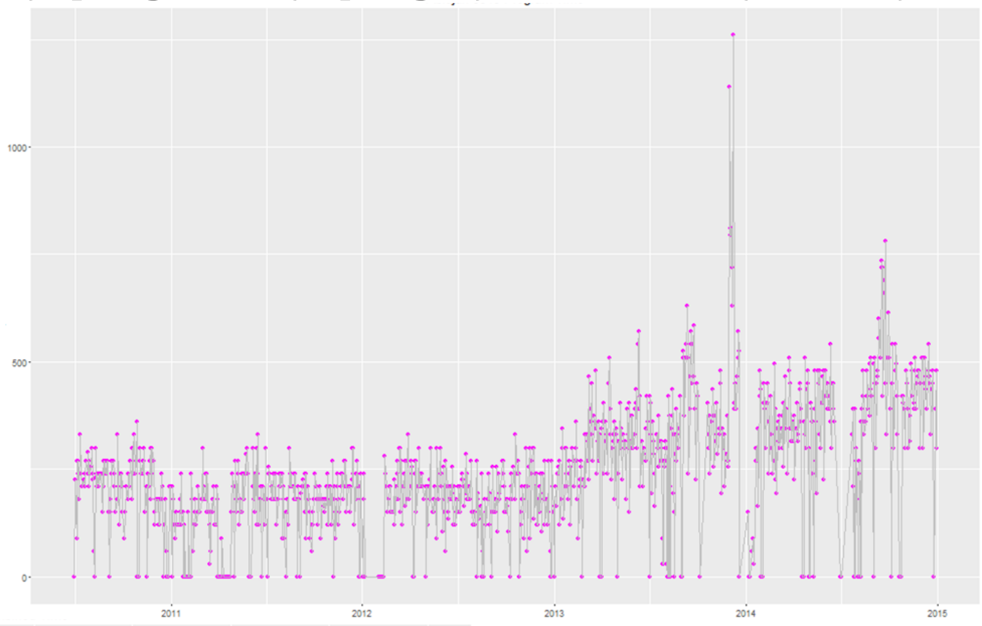
Does this represent “valid” or “invalid” data?

Data Reliability
Employee-Level Dataset Review

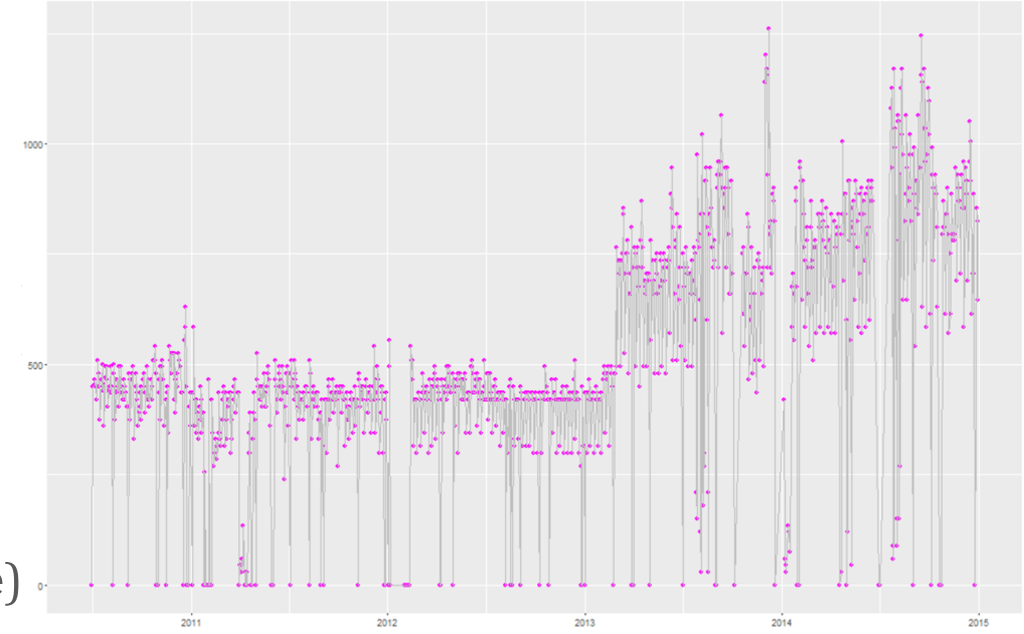
- Although we prefer to study employee time series within the **mission context**, or within the universe of **all employee** time series, individual time series can also be studied in isolation.
  - As an illustration consider a randomly selected employee, for whom the time spent on **cases**, **program activities**, and **combined** cases and program activities are shown on the next slide.
  - An interesting feature of these graphs is that the employee's hours sees a **trend shift** early in 2013; without domain expertise on the actual situation, there are multiple possible interpretations: notable ones include the employee **switching from part-time to full-time**, or a **shift in the number** of cases, services, and program activities for that mission.

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

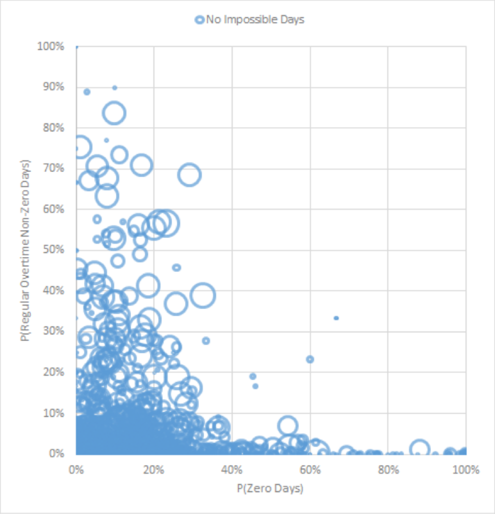


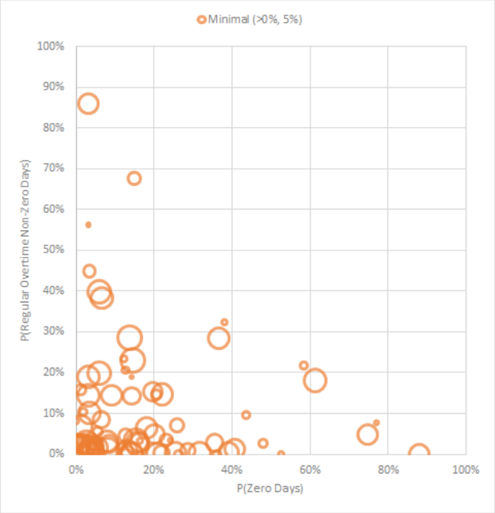


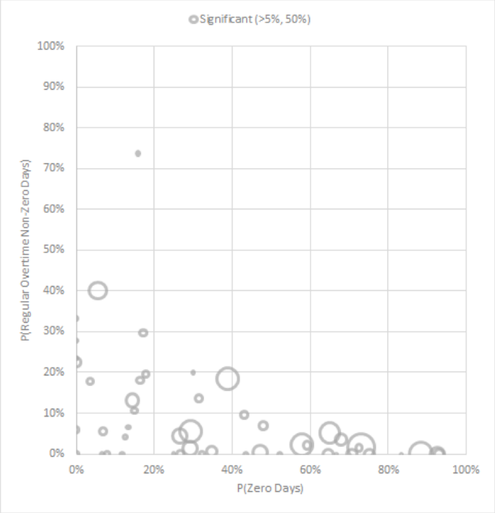
(randomly selected employee)

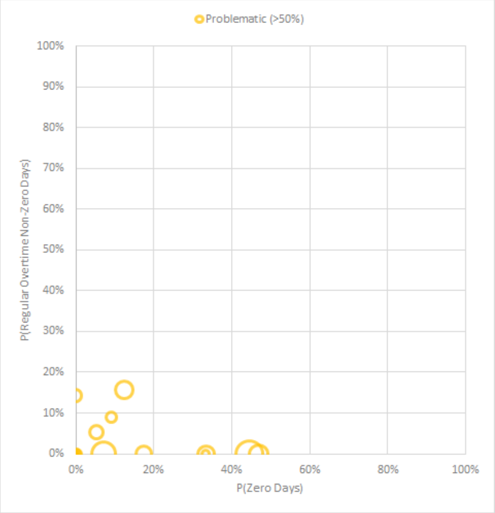


Proportion of Anomalous Overtime Days against Proportion of Plausible Days, by employee and Proportion of Impossible Days (bubble size is linked to number of entries per employee).

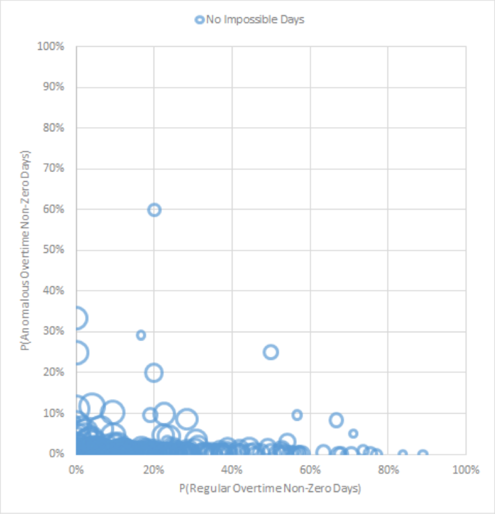


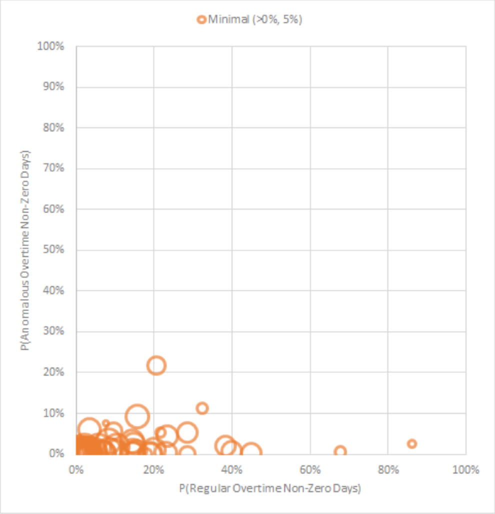


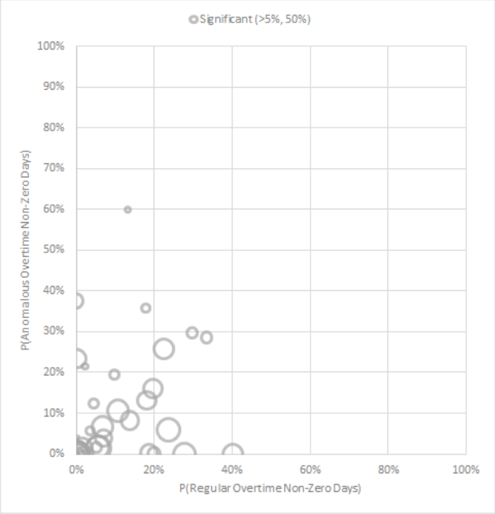


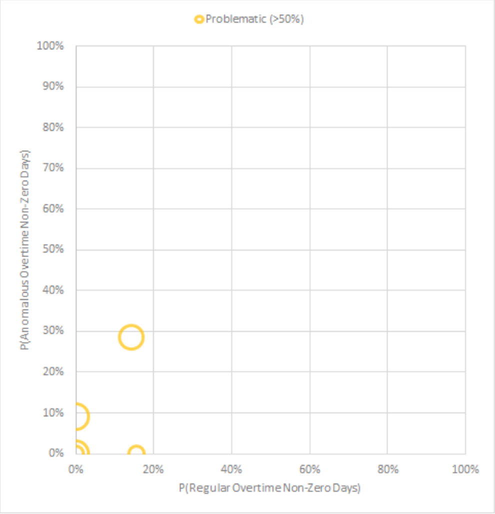


Proportion of Zero Days against Proportion of Anomalous Overtime Days, by employee and Proportion of Impossible Days (bubble size is linked to number of entries per employee).









Data Reliability
Data Entry Scenarios

- Another potential strategy for detection of invalid data is the **generation of hypothetical data entry patterns** based on a **consideration of possible data entry strategies**.

- This requires some understanding of data entry behaviours, as well as the **potential underlying causes** of invalid data entry by employees. General possibilities for this include:
  - Employees are **inaccurately remembering** what actually occurred at the mission over a given amount of time;
  - Employees are entering the data so that it **reflects a certain desired reality**, as opposed to what is actually occurring;
  - Employees wish to **comply with the requirement to enter data**, but don't know what the correct data is, so they generate data randomly and enter it into the system.

- These three broad categories can be further broken down into a large number of **specific scenarios**.
  - Suggested tests to detect these are provided whenever available.
  - **There is very little evidence to suggest that any of the candidates found in the datasets actually correspond to the scenarios they are meant to illustrate**: without external information and domain expertise, legitimate time series may look suspicious or anomalous.

Data Reliability
Data Entry Scenarios

- **Scenario 1**: daily working hours on each task are entered by computing an average from the monthly estimated totals.
  - **Results**: daily working hours should be evenly distributed for long stretches of time.
  - **Challenges**: no auxiliary (auditing) information exists to differentiate from legitimate series.
- **Scenario 2**: daily working hours on each task are entered using a reasonable guess based on past patterns and/or memory.
  - **Results**: daily working hours are not likely to be evenly distributed.
  - **Challenges**: no auxiliary information can be used to verify whether this happened since we do not know how many case/services/programs they have been working on (only the number of new open cases/services is available).
  - **Methods**: time series analysis and Benford's law could be used, but they are unlikely to yield anything useful due to the above-mentioned challenge and the fact that the reported times are unlikely to be random in the Benford sense (the daily working hours for many employees is likely to be centered around 7.5 or 8 hours, say).

Data Reliability
Data Entry Scenarios

- **Scenario 3**: daily working hours on each task are entered with typos.
  - **Results**: daily working hours should be evenly distributed for long stretches of time.
  - **Methods**: check the consistency of the time formatting; generate the distributions of daily working hours for each mission, etc.
- **Scenario 4**: daily working hours on each task are randomly entered, independently on the actual time spent on cases, services, and/or programs.
  - **Challenges**: could be difficult to differentiate from an employee whose tasks provide for rather random times.
  - **Methods**: Benford's Law might flag some anomalies.
- **Scenario 5**: daily working hours on each task are entered by copying another employee's time sheet.
  - **Results**: daily working hours for two (or more employees) are identical for long stretches of time.
  - **Challenges**: it's quite conceivable that two employees have similar responsibilities, and that sequences of matching times are not indicative of invalid data.
  - **Methods**: comparisons of matching subsets of a mission's employee's daily reports.

Data Reliability
Data Entry Scenarios

- **Scenario 6**: daily working hours on each task are zero, or near zero.
  - **Results**: time series plots will show a large number of zeros or near-zero values.
  - **Challenges**: there could be many reasons why zeros could appear.
  - **Methods**: look at distributions of employees' reported daily working times and seek distributions with large numbers of zero or near-zero values.
- **Scenario 7**: overtime daily working hours on each task for a busy week are distributed during the following week.
  - **Results**: looking at an employee's time series plot of daily working hours, the times when we expected to see a jump turn out to be flat or lower.
  - **Challenges**: without external data against which to validate this, it is nearly impossible to determine if data is naturally flat, or if overtime has been spread to subsequent dates.
- **Scenario 8**: daily working hours on each task are entered to another employee's code.
  - **Results**: some employees will be under-represented in the mission, whereas others will be over-represented.
  - **Challenges**: without auditing data, this could look the same as a busy employee and a part-time employee.

Data Reliability
Data Entry Scenarios

- **Scenario 9**: daily working hours on each task are inflated before being entered.
  - **Results**: an employee's records will be inflated.
  - **Challenges**: since the mission-to-mission data is all over the place, we would need auditing data to compare an employee's records with actual working times.
  - **Methods**: if values and patterns tend to be similar from mission-to-mission, we can compare an employee's time series with an average and flag it if it looks abnormal.

- **Scenario 10**: an experienced employee is replaced by a less experienced employee, or an inexperienced employee becomes more proficient.
  - **Results**: this could potentially show as an increase (or decrease) in daily working hours for the same tasks.
  - **Challenges**: the new employee might be as competent as the old one (or even more competent), so we could see a decrease instead, but in either case, an increase/decrease in reported times could also be associated with an increase/decrease in the number of specific cases/services/programs.

Data Reliability
Data Entry Scenarios

- **Scenario 11**: the nature of the relationship between Canada and the mission's Host Country has changed, or dramatic events are occurring in the Host Country.
  - **Results**: this could show as an increase (or decrease) in the time series for all employees if the number of employees stays constant.
  - **Challenges**: mathematical techniques cannot guess at changing geo-political relationships; without expert knowledge of the situation, it could be difficult to differentiate such scenarios from a change in employment status.
- **Scenario 12**: A mission's number of employees changed, affecting the daily working hours on each task.
  - **Results**: this could change the overall pattern of daily working hours for the mission (assuming that the change in employees was not driven by a decrease or increase in cases/services/programs, which could in turn affect any future analysis). This is fairly easy to identify visually.

Recommendations

Recommendations

- It has already been noted that it is **inherently difficult** to validate mission data.
- Some measures can be undertaken to improve the **overall validity** (over time), such as:
  - **including a list of mission-specific services** to add another basic check against invalid data by making sure no time is connected to services that are not offered;
  - **including a list of historical mission-specific events** that could affect the nature and trends in the reported data (such as new trade agreements, humanitarian crises, etc.);
  - **performing audits** (in-house verification) to determine what proportion of the reports are inaccurate, on average;
  - **changing the reporting standards/guidelines**;
  - **tracking case work and service instances to specific employees** as the work is being done, to provide average times spent on case and service instances (by mission and/or employee) which can be used to flag anomalous time reports for further examination, and
  - **eliminating the need to report to the nearest minute** as the data supports reporting to the nearest 5, 10, or 15 minutes –this could help eliminate some of the typos that were observed in the data.

Consulting Post-Mortem

Consulting Post-Mortem

- Project **did not start** as a data reliability project
  - initial analysis showed that there were some data validity issues
  - consulting team was split: contract called for data analysis, not data reliability assessment
- Client had an **unofficial objective** in mind for the project
  - there was a desire to overhaul the data collection system, but it cost the department a fair amount and it was hoped that we would find some evidence to cast doubt on the validity of the data, which could provide some impetus for an updated or new system
  - client was not prepared for the overwhelming support we found for overhauling the system, especially since reports had been released using the "invalid" data
  - client agreed to shift scope from analysis to reliability assessment as any analysis would have been based on faulty data
- Client believed in the importance of data-driven decision support, but **did not have a plan to use the data to that effect** once the collection system had been overhauled
  - a second consulting project was set up to provide suggestions
- There were some issues as the data visualisation **flagged specific employees** as habitual producer of "invalid" self-reports
  - data is not just bits on the cloud – it represents people who may have legitimate grievances at being singled out
  - consultants may identify individuals or mission to investigate further, but we rarely have the full picture at our disposal