Assessment and Analysis of BFO Consular Data

Adèle Weyl, Martin Kerdaniel, Mala Hasselink
Turnstile Analytics

Nov 01, 2021

Contents

List of Figures

List of Tables

1 Project Overview and Statement

Within Borealian Foreign Office (BFO), Consular Affairs (CA) has a software application (SPACE, produced by OneEarth) that tracks consular activity statistics and notes, with a focus on case management. Broadly speaking, SPACE is used to enable consulates to provide assistance to their consular clients while at the same time helping to identify where the workload stresses are and to provide basic statistics for requests from journalists and others.

While originally designed for client support, there is a wealth of information in SPACE databases that could potentially be utilized to the advantage of BFO. More specifically, PIMENTO (a module of SPACE), tracks the time required by employees to perform consular tasks abroad. 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 is used to evaluate the need for resources in missions. It is in fact the pivotal element when determining whether to staff, delete, or create positions. The software is scheduled to be updated/replaced in late 2022, and BFO are looking for an opportunity to determine if the current system meets their needs, and what changes should be implemented to improve its effectiveness.

1.1 Report Objectives

In order to support decisions relating to consular resource allocation and effectiveness more broadly, the SPACE data must, at a minimum, have a sufficiently high level of validity to allow reliable conclusions to be drawn from the data. Related to this, CA would like to know, and this report will consider:

- to what extent the reliablity and accuracy of the current dataset can be verified, especially given that it is input by users without extensive checks possible, and
- if, given its current level of validity, the existing dataset can be reasonably used for the purposes and types of decision making described above.

The results of this analysis are described in detail in Section 3.

Apart from issues of data validity, data can only be used to support particular types of decisions if it can be suitably connected to these decisions. Thus CA would also like to know if the type of data currently being collected can be reasonably and effectively used for the desired decision making purposes. To answer this question, we further assessed the existing data to determine the ways in which it could be used to meet CA's requirements to monitor the delivery of the consular program. The results of this assessment are described in detail in Section 4.

Finally, CA would like to know what additional data (if any) would be required to improve their decision making capabilities, and minimize inaccuracies that may have been identified in their data collection. To answer this question we have suggested modifications to CA's data collection strategy, and recommended ways to improve their analysis using existing and other tools. These recommendations are provided at the end of each of the main sections of the report.

1.2 Key Points and Recommendations

Although the report will go into further depth on all of these points, key recommendations coming out of the examination of the consular data are as follows:

- Due to the nature of consular work (in particular, the variability of the work from mission to mission and month to month) it is inherently difficult to determine the reliability and validity of the mission level data using only measures internal to the data. It may be possible, however, to provide some partial measures which can be combined with external information to provide some ability to judge the reliability of mission data in particular instances. For further discussion, see Section 3.
- As illustrated by the time series provided in Section 3 and further discussed in Section 4, this variability across missions must also be properly taken into account during the calculation of any performance metrics, with missions only compared to other demonstrably similar missions (e.g. using techniques like clustering). Otherwise, these metrics will not accurately reflect mission performance.
- The presence of separately entered daily and monthly work logs has some provisional advantage with respect to determining the facility with which missions are using available data entry systems. For further discussion, see Appendix B. However, this advantage is entirely outweighed by the data interpretation challenges that results from this data collection system. Thus it is recommended that the daily and monthly logs be synchronized.
- In several cases the consular data that is collected (or can be derived from other data that is collected) is partial, or incomplete (see Tables 1 to 3 for more information on this). In such cases, it may be tempting to use this incomplete data as if it were comprehensive, but doing this will most likely yield false conclusions. Thus, despite its availability, partial data should not be used.
- In effect, the existence of this partial data results in the illusion of having more data than is actually available. A good goal for future system redesign is to make changes such that data be collected in a comprehensive manner, so that it can be properly used for decision support. As it stands, it is difficult to calculate valid performance metrics with the available data.
- In terms of future data collection, our most important recommendation is to begin to track the start and end times (and by extension duration) of case and service activities, as well as the employees who are working on a given case or service at any point in time. Without this data, a variety of measures that are essential for determining data validity, resource sufficiency and mission efficiency cannot be calculated. For further discussion, see Sections 3 and 4.
- If systems are, first, redesigned to fully capture the required data and then combined with existing external data, there is the potential to calculate a number of useful mission-level metrics, as well as, more generally, a mission snap-shot that can provide an at-a-glance summary of relevant mission level information. Both of these system outputs could then be used to facilitate decision making relating to mission management. For further discussion, see Section 4.

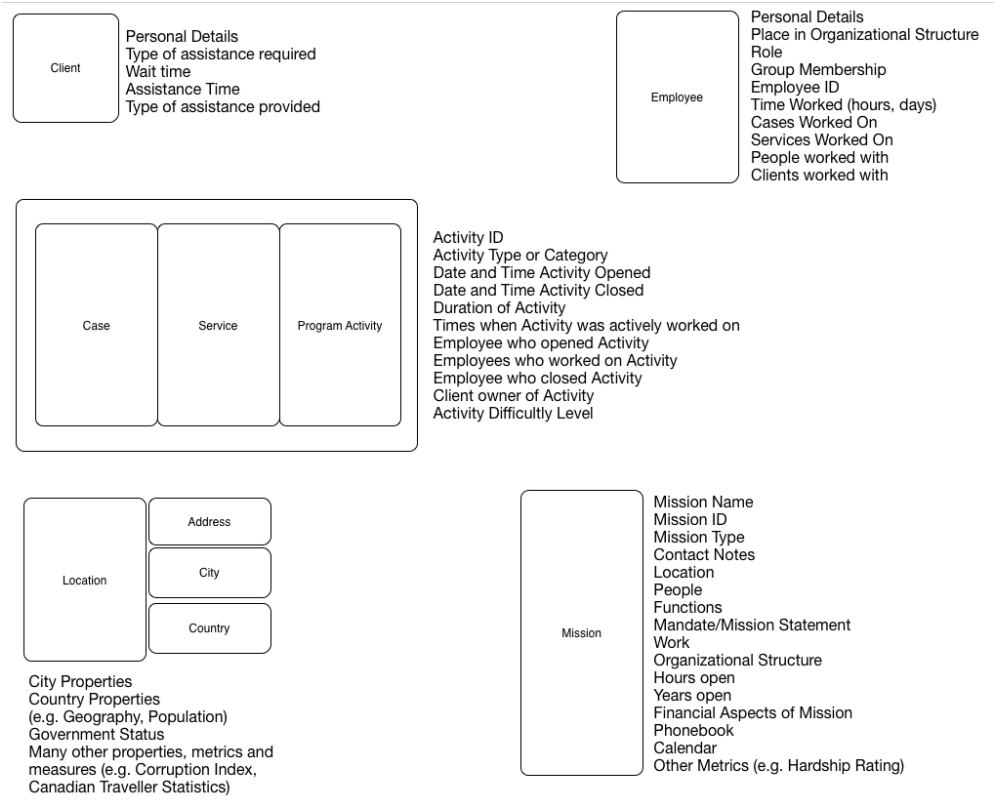


Figure 1: Key objects in the consular network and some properties of these objects.

2 Description of the Consular System and Data

In order to assess the data provided by the SPACE system, both in terms of its validity and its decision support suitability, it was necessary to develop a broad understanding of the consular system itself. This understanding was built first through an examination of the structure of the supplied dataset, and then, following this, through interviews with CA.

2.1 Consular System Description

This review resulted in the development of a picture of the consular system, including key objects and relevant object properties. The available data was then compared to this picture in order to understand how comprehensive the dataset was, as well as in order to identify areas which might be vulnerable to data issues and areas where additional data collection might usefully augment the current data collection strategy.

Figure 1 depicts the key objects and some of the objects properties that were identified. For a more detailed list of relevant consular system objects and properties, see Tables 1 to 3.

2.2 Data Set Description

Our current understanding of the SPACE dataset suggests that the data of primary interest for consular management is contained in four of the PIMENTO tables – specifically, those which provide

Object Properties	Data Availability
Functions	Fully Available
Types of Services Provided	Fully Available
Types of Cases Provided	Fully Available
Types of Consular Programs	Fully Available
Connected Objects	
Missions	

(a) Consular Network

Object Properties	Data Availability
[Mission ID]	Auto-generate
Mission Name	Fully Available
Mission Type	Fully Available
Contact Notes	Fully Available
Mission Profile	Fully Available
Functions	Fully Available
Mandate/Mission Statement	Potentially Available
Hours Open	Fully Available
Years Open	Derivable
Financial Aspects of Mission	Potentially Available
Connected Objects	
Building	
City	
Country	
Mission Employees	
Mission Clients	
Local Service Providers	
Contacts	
Bureaucractic Liasons	
Mission Work	
Organizational Structure	

(b) Missions

Table 1: Further details of relevant objects and object properties in the consular network – I.

logs of mission activities (cases, services and programs) as well as time spent on these mission activities, for each day and also for each month. Within these tables, data is available across a time span of 10 years, from 2011 – 2020.

The system was upgraded in 2016. During the upgrade, the categories relating to cases and services were changed, resulting in a break in the dataset at this time. Discussions with CA also confirmed that, as result of the upgrade, data accuracy should be highest from mid – 2016 onwards. For this report, a subset of data from these four tables was reviewed in depth. Specifically, the focus of this report is on an analysis of case, service and program related data collected between July 2016 and December 2020.

Tables 1 to 3 note which key consular object properties either correspond to, or can be derived from, one or more of the fields in these four PIMENTO tables. The possible availability of additional data will be discussed further in Section 4.

Object Properties	Data Availability
Time Worked	Partially Available
Work Difficulty Assessment	Possible Metric
Types of Services Provided by Mission	Partially Available
Types of Cases Provided By Mission	Partially Available
Types of Programs Worked on by Mission	Partially Available
Tally of Cases of Each Type Opened on a Given Date	Fully Available
Tally of Services of Each Type Provided on a Given Date	Fully Available
Hours Worked on Program Activities of a Given Type on a Given Date	Fully Available
Connected Objects	
Service	
Case	
Program Activity	

(a) Mission Work

Object Properties	Data Availability
Reporting/Responsibility Structure	Potentially Available
(Roles)	Potentially Available
Role Title	Potentially Available
Role Responsibilities	Potentially Available
Role Activities	Potentially Available
Reports to	Potentially Available
Supervises	Potentially Available
(Groups)	Potentially Available
Group Title	Potentially Available

(b) Organizational Structure

Object Properties	Data Availability
Contact Details	Potentially Available
Place in Organizational Structure	Potentially Available
Role	Potentially Available
Group Membership	Potentially Available
Employee ID	Potentially Available
Time Worked (hours, days)	Potentially Available
Cases Worked On	Partially Available
Services Worked On	Partially Available
Local Employees Worked With	Not Available
Clients Worked With	Partially Available
Work Experience	Potentially Available

(c) Mission Employees

Object Properties	Data Availability
[Client ID]	Auto-generate
Contact Information	Potentially Available
Reason for Contacting Mission	Not Currently Available
Wait Time (before first seeing someone)	Not Currently Available
Duration of Time (before conclusions of interaction with m	Not Currently Available
Type of Assistance Provided	Not Currently Available
Client Satisfaction (with experience and assistance provide	Not Currently Available

(d) Mission Clients

Table 2: Further details of relevant objects and object properties in the consular network – II.

Object Properties	Data Availability
Mission Employee Worked With	Not Currently Available
Dates	Not Currently Available
Type of Assistance Provided	Not Currently Available

(a) Service Providers

Object Properties	Data Availability
(Building + City)	Potentially Available
Address	Potentially Available
Building Status	Potentially Available
Occupancy Type	Potentially Available
Ease with which people can get to building	Potentially Available
Security	Potentially Available
(Country)	Potentially Available
Geography	Potentially Available
Government	Potentially Available
Visa Requirements	Potentially Available
Many other properties, metrics and measures	Potentially Available

(b) Location

Object Properties	Data Availability
Case ID	Auto-generate
Case Type	Potentially Available
Date and Time Case Opened	Potentially Available
Date and Time Case Closed	Not Currently Available
Duration of Case	Not Currently Available
Active Work Times	Not Currently Available
People who have Worked on Case	Not Currently Available
Client Owner of Case	Not Currently Available
Case Difficultly Level	Possible Metric

(c) Cases

Object Properties	Data Availability
Service Provision ID	Auto-generate
Service Type	Not Currently Available
Duration of Service Provision	Not Currently Available
Date of Service Provision	Not Currently Available
Employees who Provided Service	Not Currently Available
Local Service People who Provided Service	Not Currently Available
Service Difficulty Level	Possible Metric
Connected with Case	Not Currently Available
Client Served	Not Currently Available

(d) Services (Instances)

Object Properties	Data Availability
Program Type	Fully Available
Hours Worked	Fully Available
Employees Involved	Fully Available
Activity Objective	Potentially Available
Date Objective Achieved	Not Currently Available

(e) Program Activities

Table 3: Further details of relevant objects and object properties in the consular network – III.

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**

Table 4: Proportion of impossible days, per mission and per employee. The vast majority of missions and employees never enter an impossible number of minutes.

3 Data Set Reliability Assessment

In any data analytical endeavour, the quality of the output is strongly affected by the quality of the input. This is singularly *à propos* for the dataset under consideration since the data is self-reported.

A note on monthly data: Based on discussions with CA, it is understood that monthly log data should, in some sense, be viewed as having more inherent validity than the daily log data, because monthly log data must be reviewed by management before being submitted into the system, and this oversight may be sufficient to ensure greater validity of that data. The daily log data by contrast may be entered less diligently for a number of reasons, not least of which being that it is currently not a requirement for the daily log data to be entered in order for a monthly log to be produced.

The challenge in viewing the daily and monthly log data in this manner, and essentially abandoning the daily log data entered into the system, is that, in reality, it is impossible to create monthly log data that in any way *accurately* reflects the reality of monthly work taking place in the mission without recourse to some information from employees about the work they did on a daily basis, over the course of the month. Thus, whether or not daily data is entered into the system or not, some kind of daily data is *de facto* being used to create the monthly logs. Consequently, while certain types of analysis of the consular data may perhaps be conducted using monthly aggregates, data validation has to occur at the daily data level; although it may turn out that the daily data as it is currently collected is sub-optimal, any validation technique must be applied and demonstrated in relation to this daily data.

3.1 Data Set Basic Checks

As a first step towards assessing the reliability of the data, basic data checks were carried out using standard mathematical and logical tools to verify the consistency and completeness of the dataset. The results of these baseline checks are included in Appendix A. The general findings from these checks were that the data was clean (e.g. fields had valid values in expected ranges), but that there were significant gaps in data entry, as well as logical inconsistencies resulting from data entries issues. Our recommendations for addressing these issues are provided at the end of this section.

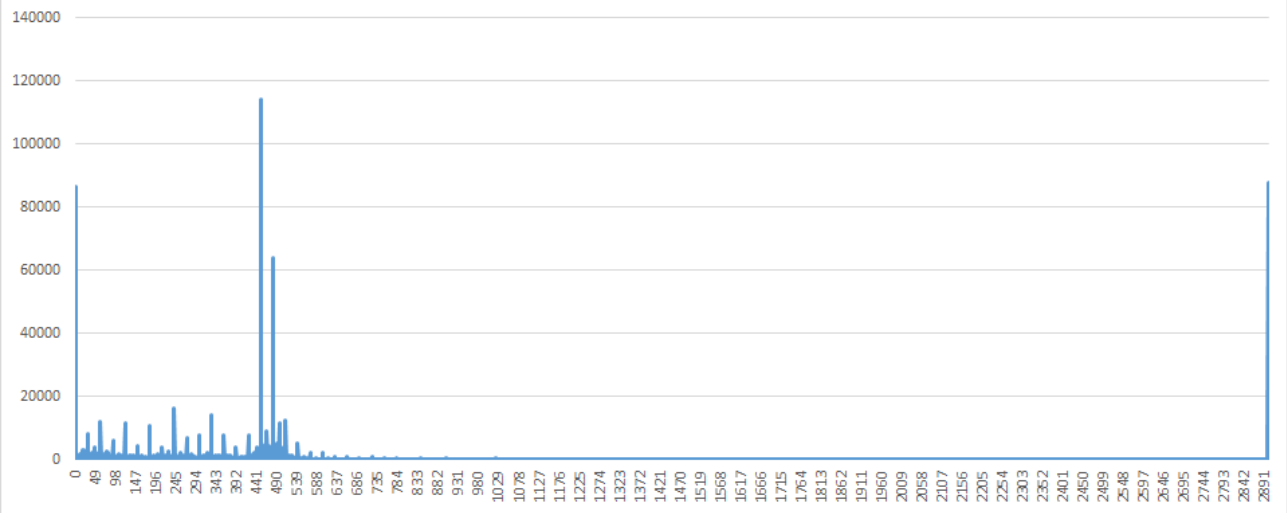


Figure 2: Frequency of reported daily work time values. Peaks at 0 min, 450 mins (7.5 hrs), and 480 mins (8 hrs). All entries greater than 2900 mins are put in the same bin (2901).

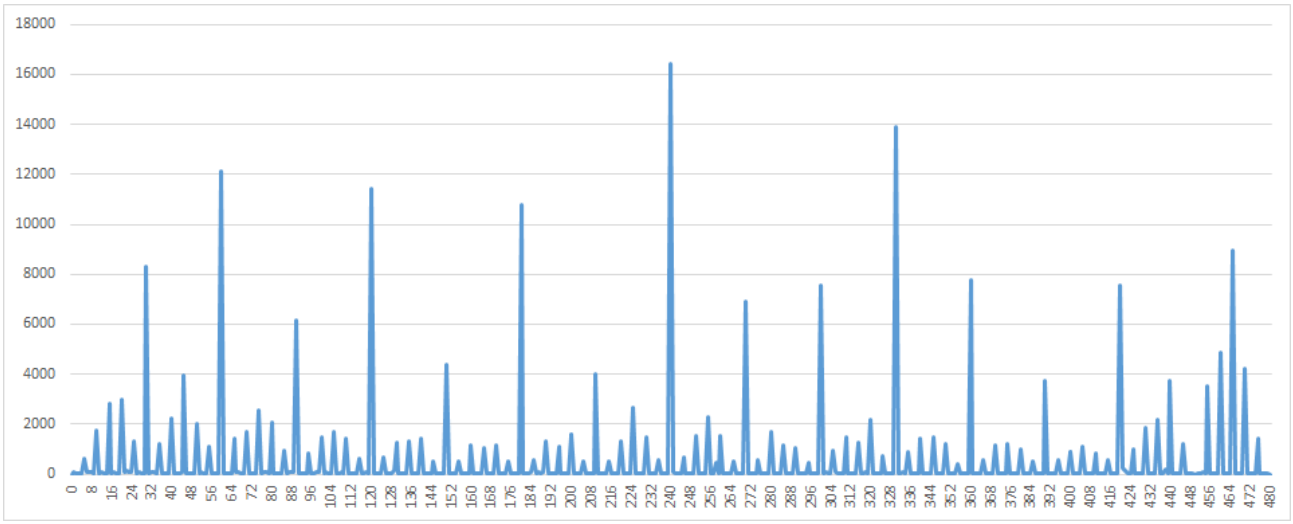
3.2 Review Over Entire Dataset

Following these basic data checks, a more in-depth analysis of the data was carried out, including a detailed review of data patterns evident within and across missions. The focus of this full dataset review was on employee hours worked. Via application of a number of techniques, the review revealed a mixture of data patterns which were either improbable or difficult to explain.

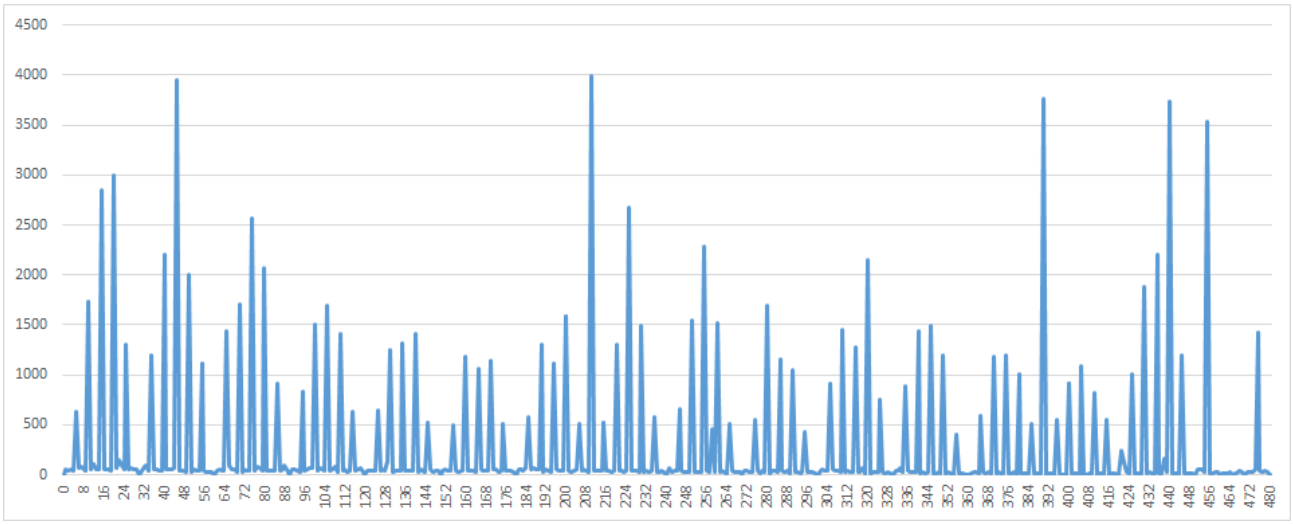
In general, time series plots can be used to detect some anomalies in data, such as recorded daily working times greater than 1440 minutes (24 hours), which may be due to multiple people's hours entered under a single employee ID, but could also be caused by data entry issues; consider Reme and Tolosa, for instance, which are very noisy due to a large number of the time entries being greater than 1440 minutes (see Figure 18, p.27). A summary of these "impossible days" at the Mission- and Employee-level is shown in Table 4.

Another basic check is to plot a bar chart of the frequency of specific daily work time values being reported by all employees (see Figure 2). Unsurprisingly, this shows that there are peaks in or around the 7.5-8 hours region (450-480 minutes), but there are also unexpected features: the number of 0 values being reported, and the number of values above 1440 mins (see Table 5).
Another issue that can be detected fairly easily is the heaping of time values: psychologically, human beings are more likely to report by rounding to the nearest 60-, 30-, 20-, 15-, 10- or 5- minute blocks. This can easily be seen in Figure 3. Here, the anticipated heaping issue 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. Do 7.5 hrs and 8 hrs appear because these are the expected number of hours to be recorded, or the actual number of worked hours?

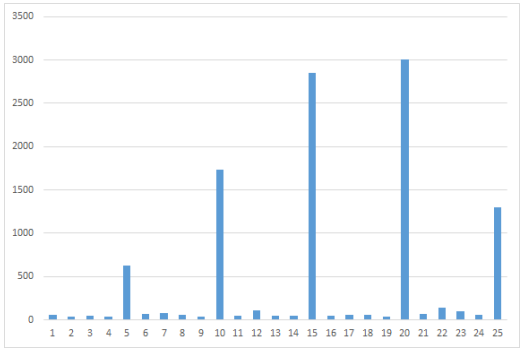
Other anomalous figures abound. Consider, for instance, the fragments of the bar chart from 341 to 344; from 924 to 936, and from 2579 to 2591 (see Figure 4). It is conceivable that an



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



(b) With counts also removed at 30-mins intervals.



(c) Over the first 25 minutes. Note the non-zero counts for non-multiples of 5 mins.

Figure 3: Heaping in the reported daily work time values. The heaping cycles can easily be seen after removal of high frequency data.

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

Table 5: Recorded daily working times (statistics).

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. For instance, employees 271 (Kingstown) and 1408 (Monte Ovidio) each reported two instances of 341 minutes, but, emblematically, their time series (while very different in nature, see Figure 5) do not allow us to differentiate between the two alternatives. The same situation holds for the other employees with similar anomalous readings: without external audit data, we fail to differentiate between valid and potentially invalid data.

Lastly, it is quite difficult to identify patterns which are universal to all missions. For instance, employees 7849 and 8305 (Davaka) seem to work from Sunday to Thursday and they have roughly the same time range (between 270 minutes and 480 minutes), with a shorter day on Thursdays; whereas in Elgin, all employees seem to record roughly the same amount on every working day (but different from one employee to the next), except for employee 5805 (see Figure 18, p.27).

Ideally, the lack of similarity between these two missions (as well as the missions in Reme and Tolosa, on the same page, or any of the other missions which have been plotted in this document) would be an indication that at least one of those contains invalid data, but since not every mission provides the same services or has, a priori, similar traffic, and since this external data is unavailable to us, there is no basic check to determine the validity of patterns (in fraud detection, compliance with Benford's Law is sometimes used, 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).

Thus, looking across the entire dataset, we find that 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 anomalous occurrences. Unfortunately, while in some circumstances these types of occurrences might indicate invalid data, in the current system, there are generally other conflating explanations which cannot be disentangled from this question of validity. This issue will be explored in further depth when looking at mission specific data. As a result, the conclusion from this across dataset review is that, although common data analysis techniques can be used to detect a variety of unusual or anomalous patterns, interpreting these findings is made challenging by the way that information is currently collected, along with the types of information collected.

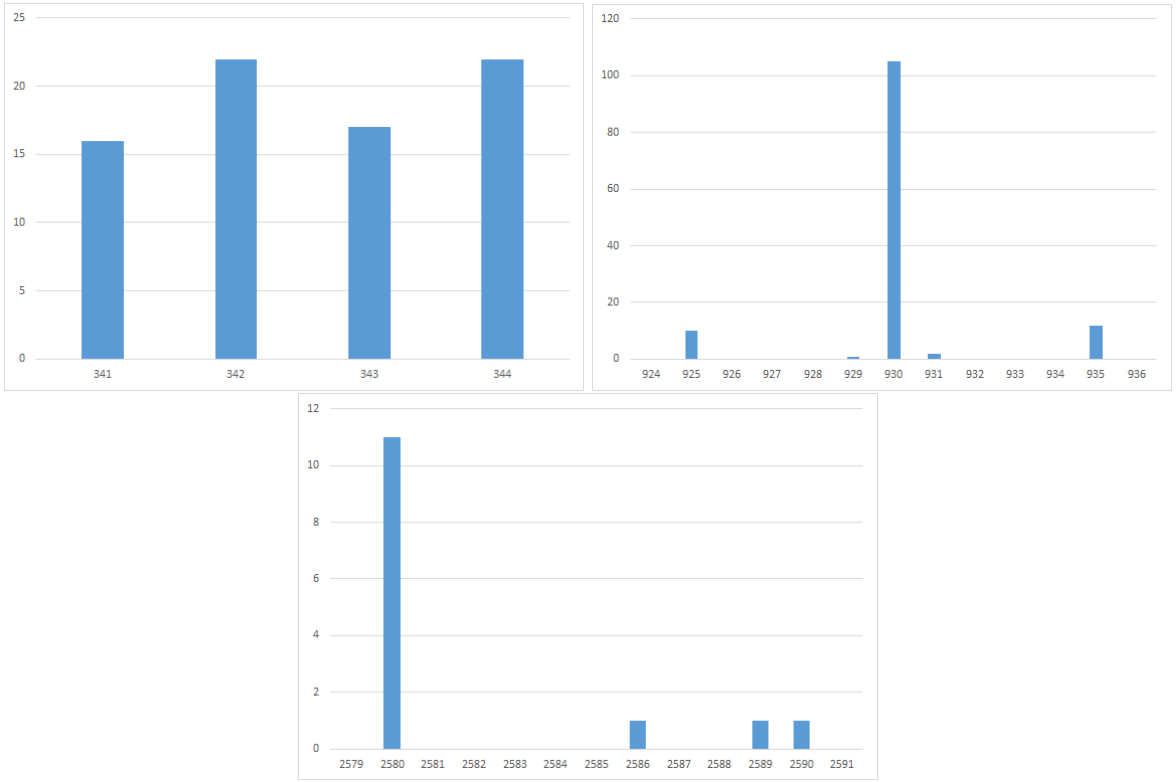


Figure 4: Reported daily work time values (extracts): 341 mins to 344 mins; 924 mins to 936 mins; 2579 mins to 2591 mins.

3.3 Mission Level Dataset Review

The question of validity of the data at the mission level is, in theory, simpler to tackle: 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. The end result of this approach would have been a data validity metric that could rate individual missions on their level of data validity, and assess the baseline integrity of the data.

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** (as in the previous section). Our hope was that the analysis of this data would provided some leads which would have helped us build the data validity metric. Ultimately, we concluded that, given the enormous variability in mission employee reports, and on mission-level aggregates, any such metric could not be considered without substantial additional external validation and auditing data. Thus, a stand-alone validity metric could not be constructed, because domain expertise about each mission (and potentially the employees) is required in order to differentiate and identify which of various interpretations is most likely to explain the observed data.

That being said, the failed attempt to construct this metric involved a number of useful observations which will be related in this section. A three pronged approach was taken in an effort

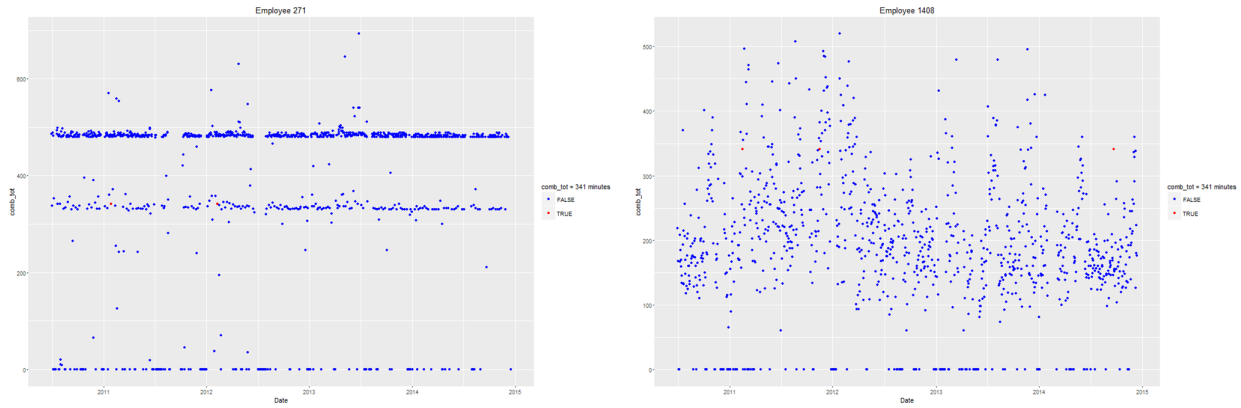


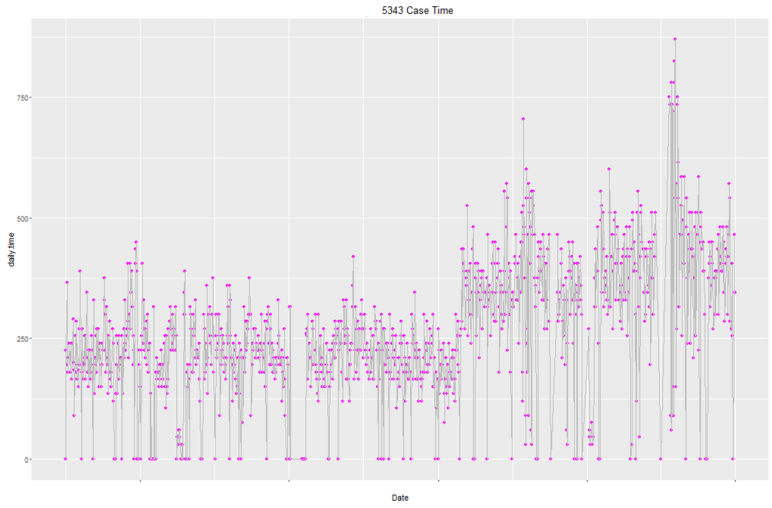
Figure 5: Time series of daily work time values for 2 employees: 271 (Kingstown), and 1408 (Monte Ovidio). Reports corresponding to 341 minutes are shown in red.

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 mission employee hours over the entire range of data (2016-2020). 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.

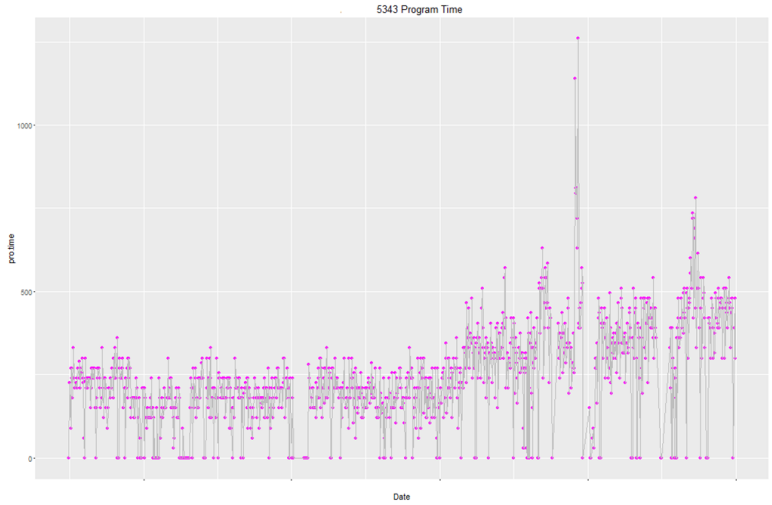
3.3.1 Employee Hours: Baseline Consistency Analysis

Discussions with CA 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. A selection of the results are provided on pages 16 to 29, which include mission time series and summaries of some salient patterns observed in the data for selected missions (the masculine gender is used neutrally).

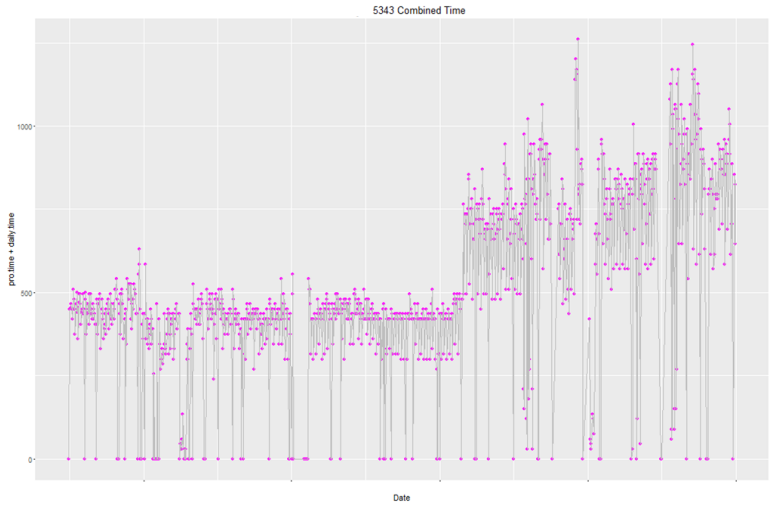
Although we prefer to study employee time series within the mission context, or as a part of the universe of all employee time series, individual employee time series can also be studied in isolation. To illustrate this approach, we randomly selected a mission (Adriata) and employee (5343) and plotted the time spent on cases, on program activities, and on combined cases and program activities (see Figure 6). An interesting feature of these graphs is that the employee's hours sees a trend shift early in 2019; 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.



(a) Time spent on cases and services.



(b) Time spent on program activities.



(c) Time spent on combined cases, services, and programs.

Figure 6: Time series for Adriata employee 5343.

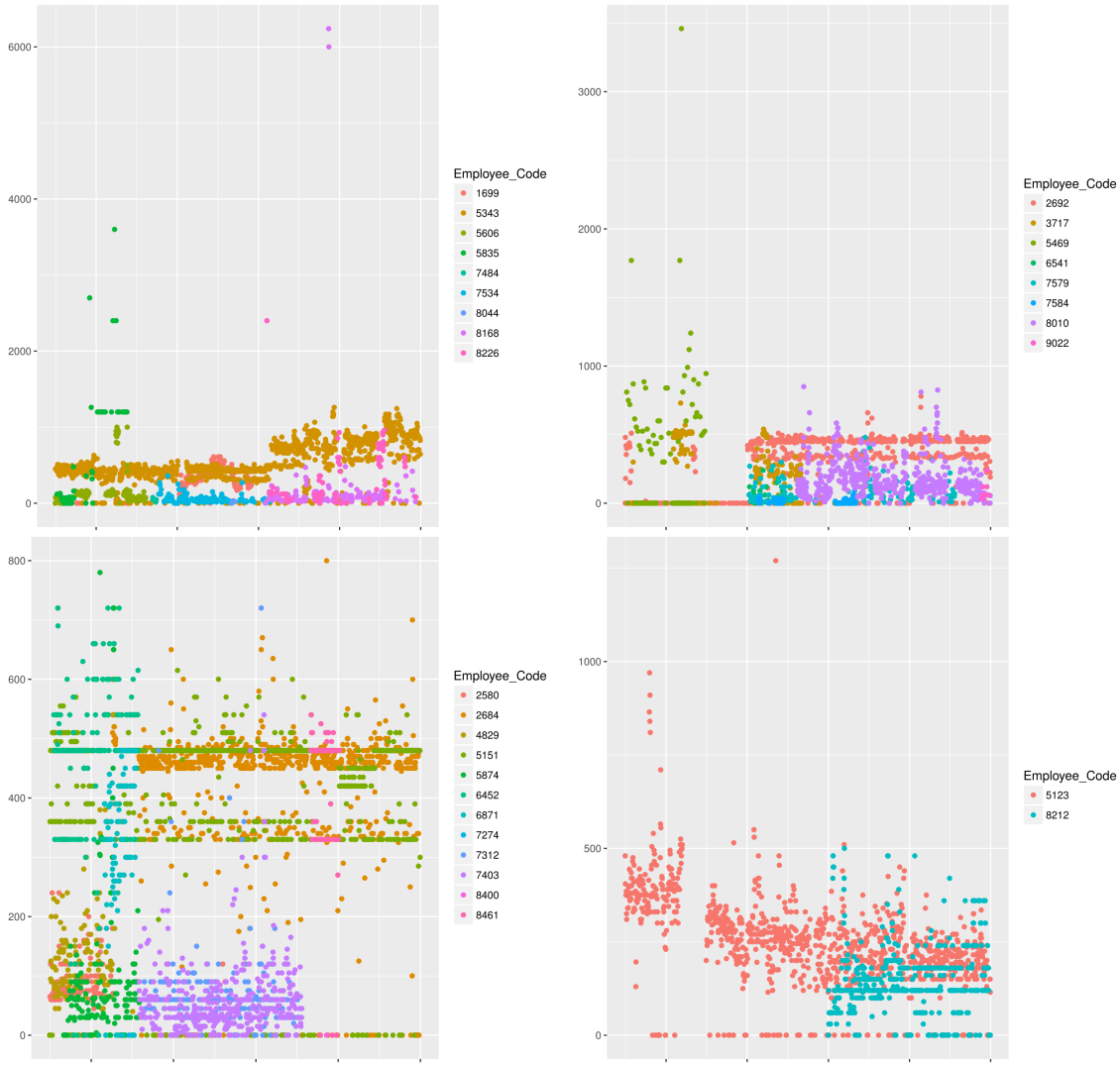


Figure 7: Daily working hours for 4 missions – I: Adriata, Addasibaba, Agruma, Aitioch.

Adriata: Employee 5343 working time increased in early 2019; 2 new employees were hired at the same time.

Addasibaba: Time series of Employee 2692 shows two clear boundaries; nobody entered time between July 2017 and January 2018.

Agruma: From July 2017 to December 2020, two employees (employee 2684 and employee 4829) tracked their times in similar ranges between 350 minutes (approximately 5.5 hours) and 500 minutes (approximately 8 hours); for the whole of 2020, only two employees worked in the mission, while there were at least three people working previously, and yet the time series remains stable.

Aitioch: Employee 5123 had a decreasing trend in his time series and his working time stabilized after another employee was hired; nobody worked between mid 2017 and 2018.

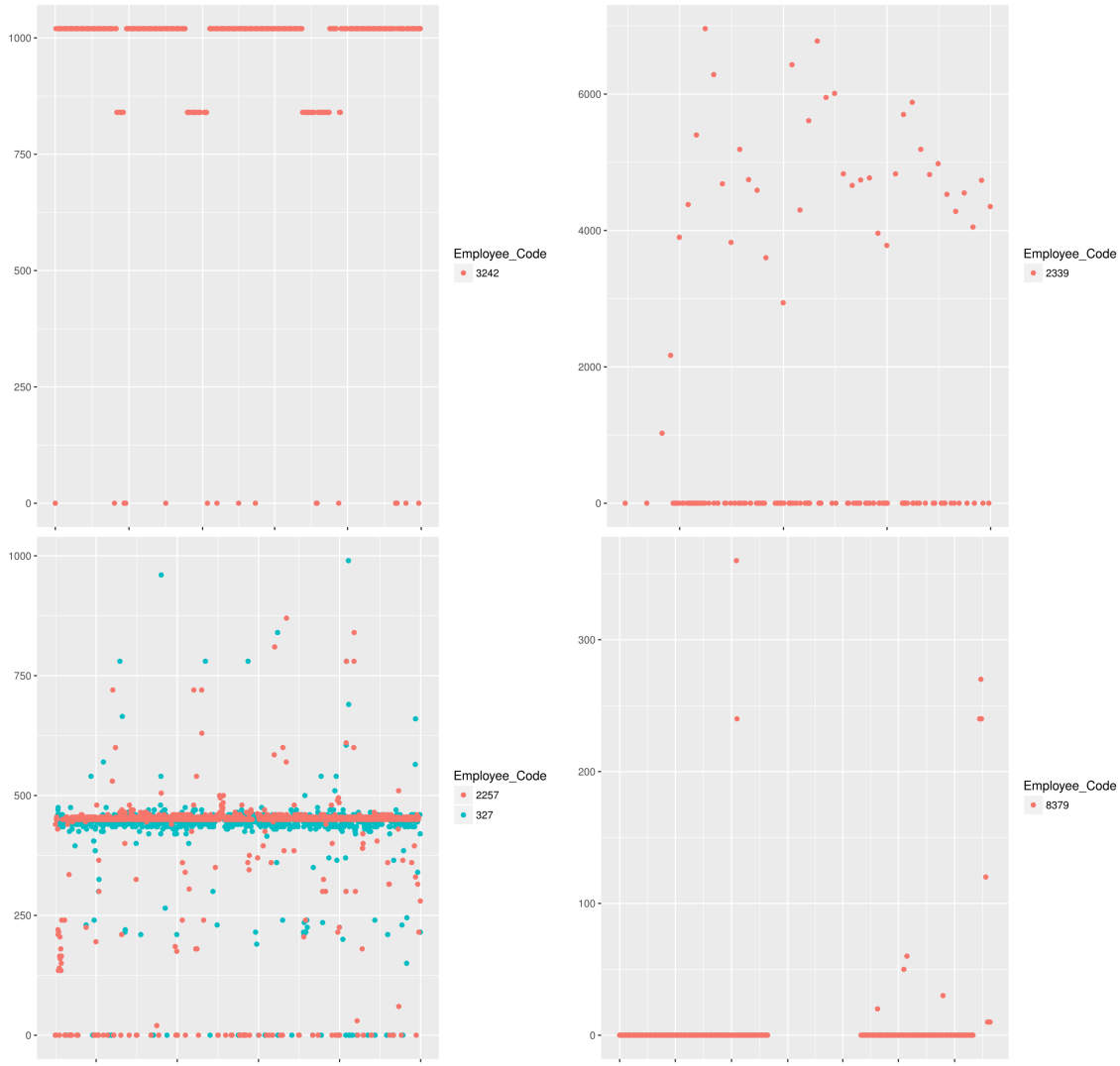


Figure 8: Daily working hours for 4 missions – II: Ahmohamabad, Al-geris, Atilanta, Bakounour.

Ahmohamabad: Time series was very stable with two clear boundaries, that are approximately 1200 minutes (20 hours) and 780 minutes (13 hours); this mission closed after October 2017.

Al-geris: Time series only contained extreme values, either greater than 1000 minutes (around 17 hours) or 0s, which could mean that the employee was not tracking his working time regularly.

Atilanta: Two employees had similar working times, around 480 minutes (8 hours) per day; the red points usually were above the blue points, which seemed to be employee 2257 working half an hour more than employee 327 on a daily basis.

Bakounour: Most of the points were 0s with only a few points having other values, generally below 400 minutes (approximately 7 hours); nobody worked from late 2019 to June 2020.

Figure 9: Daily working hours for 4 missions – III: Baotou, Bruocsella, Capeton, Casa Branca.

Baotou: Most of the working times were below 500 minutes (approximately 8 hours); one observation was above 4000 minutes (approximately 67 hours), which is unrealistic.

Bruocsella: Most of the working times were recorded around 0; a small variation occurred from July 2016 to December 2016, and from July 2020 to December 2020.

Capeton: Both of the employees have a similar lower boundary (240 minutes); the time series for employee 7336 jumped a little during the time he was working alone, from January 2019 to February 2020.

Casa Branca: The working times for employee 933 were above the other employees and showed a slight decreasing pattern after 2017; during the time he was gone, another employee was brought in to work in the mission, with most of his working times during this period above the 480 minute boundary; employee 7259 always reports exactly 0 or exactly 480 minutes.

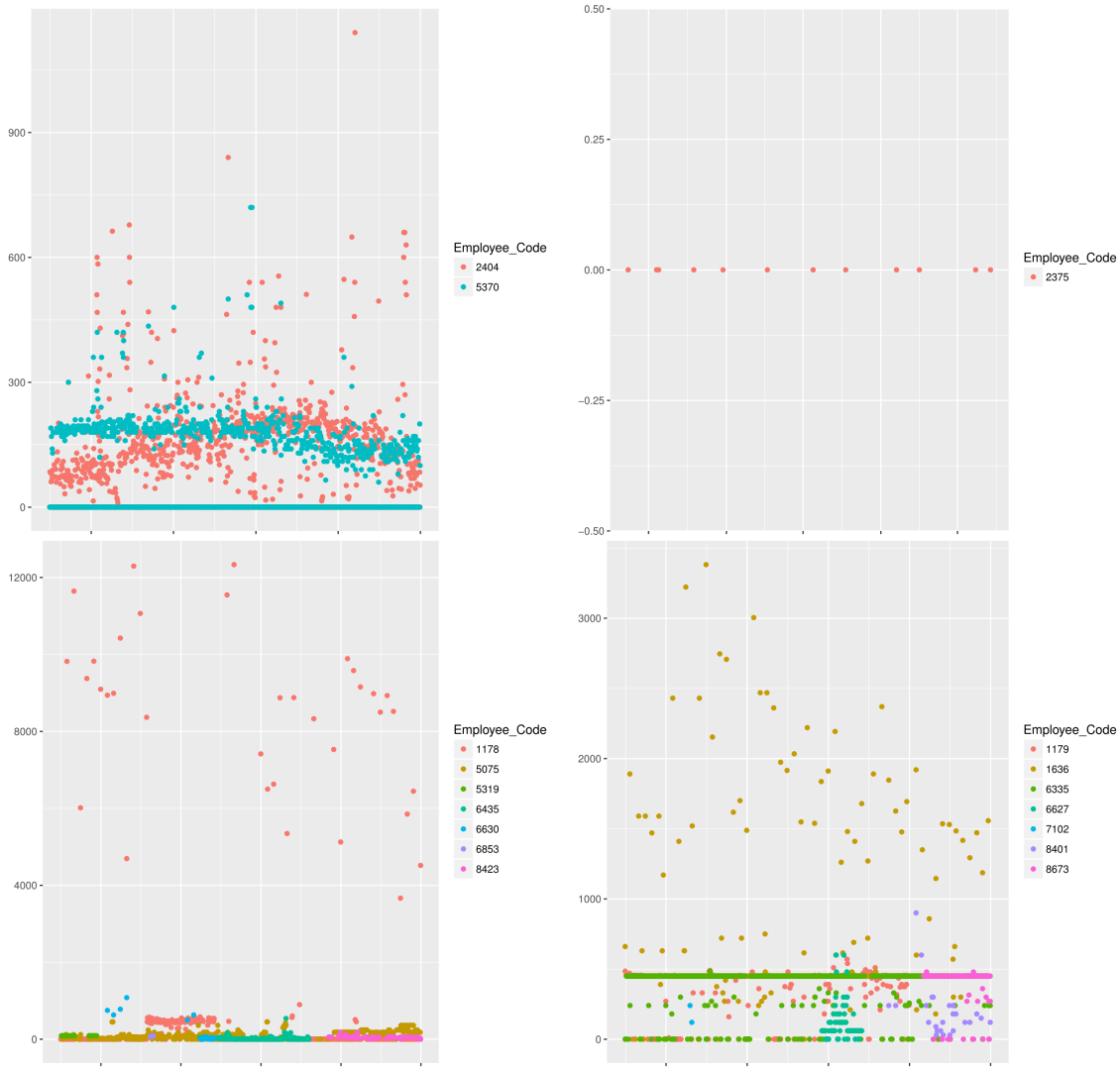


Figure 10: Daily working hours for 4 missions – IV: Copenhavn, Copper Springs, Corona, Dulles.

Copenhavn: Red points showed a growth from July 2016 to July 2019 and then decreased slightly; blue points started higher than the red ones and remained stable until January 2019; two time series displayed a similar decreasing trend after July 2019.

Copper Springs: Very sparse entries for over three years; working time showed as 1 minute for existing entries.

Corona: Employee 1178 entered extreme daily working times, but his time series showed a stable period between July 2017 to July 2018; the other employees in this mission recorded their working times as being close to 0.

Dulles: The time series of employees 6335 and 8673 were very straight; employee 1179 showed a similar pattern with little variation in the working times; generally speaking, employee 1636 had extreme working times once a month.

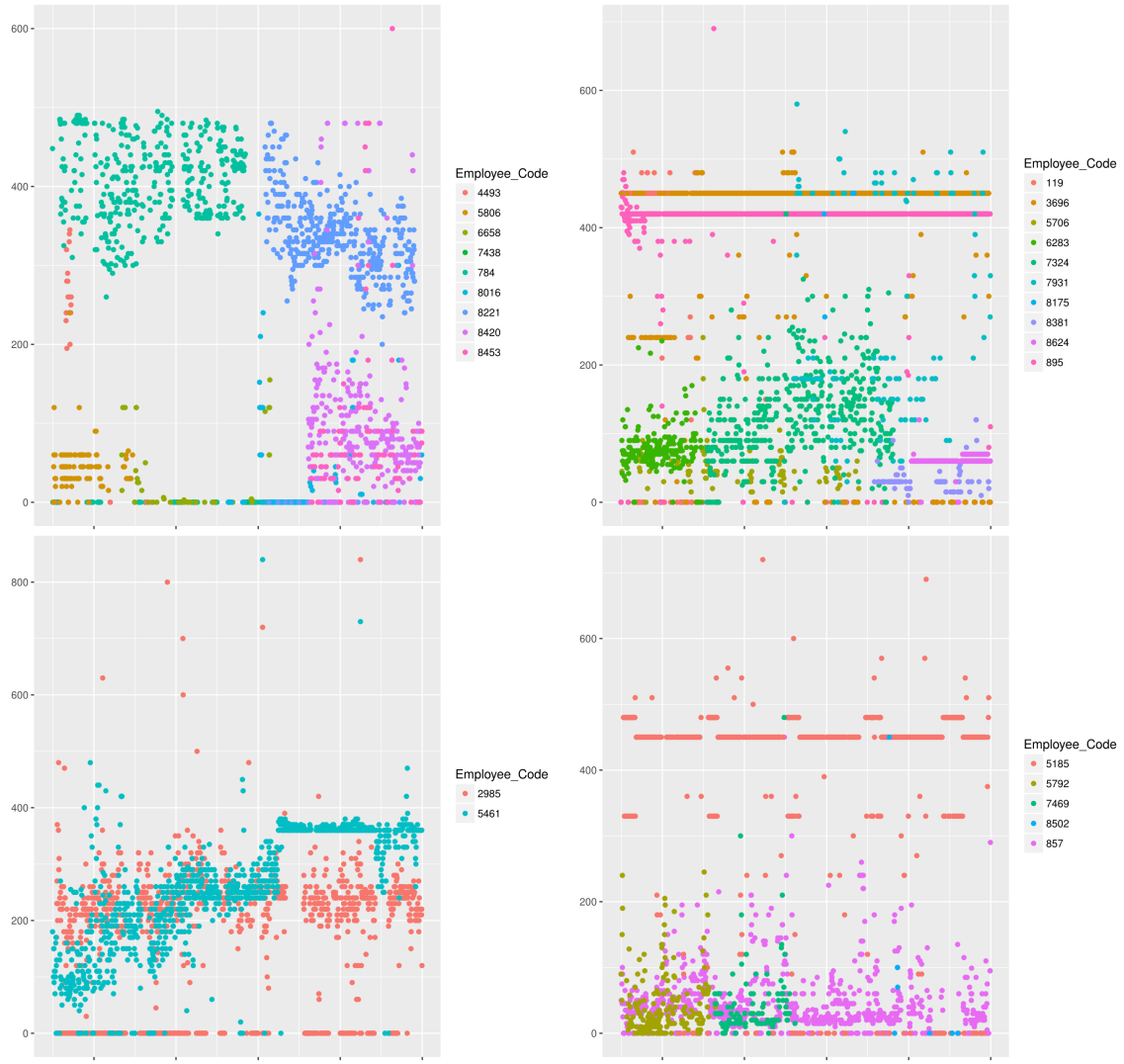


Figure 11: Daily working hours for 4 missions – V: Damascia, Douala, Guardala Ciudad, Helsingfors.

Damascia: The mission hired four new employees after an experienced employee left; the working times of employee 8221 dropped when this occurred.

Douala: Employee 3696 had a straight line for most of his working time; four employees worked under 200 minutes (approximately 3 hours) per day.

Guardala Ciudad: The same two employees worked for this mission from July 2016 onwards; most times were below 300 minutes (5 hrs) per day; employee 5461 jumped suddenly in early 2019 and remained at a high level; employee 2985 did not have much variation.

Helsingfors: Employee 5185 had an annual pattern for the time worked at this mission where July and August seemed to be busier; no such pattern was shown for the other employees; prior to July 2018, there were three employees working constantly, but the time series for employees who stayed in the mission did not vary much after July 2018.

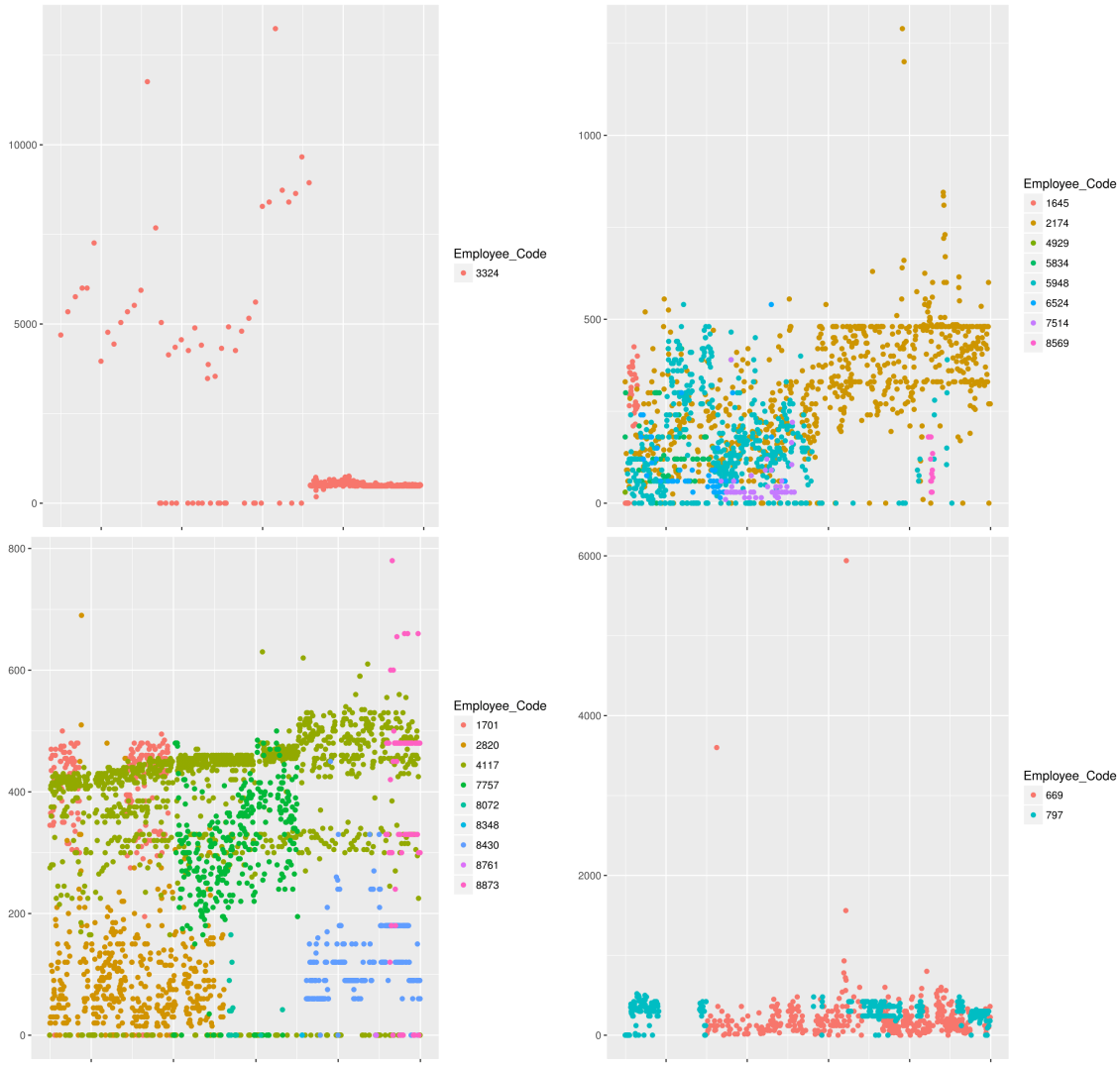


Figure 12: Daily working hours for 4 missions – VI: Kiracha, Kaartoom, Kowloon, Limonia.

Kiracha: The time series was very sparse before July 2019; the working times stabilized after that date and showed little variation.

Kaartoom: Employee 2174 showed a growth in his time series when there were fewer employees working in the mission (two lines after December 2018); time entries for employees 5948 and 8569 were limited after that date.

Kowloon: The time series of employee 4117 went up steadily from July 2016 to July 2019 but started varying more after a new employee was brought in (July 2019 to December 2020); the new employee 8430 had lower values for his working time than the other employees.

Limonia: Employee 669 worked constantly from July 2017 to December 2020, while employee 797 did not; nobody worked from December 2016 to June 2017; some anomalous observations can be found.

Figure 13: Daily working hours for 4 missions – VII: Llangollen, Lochemburg, Mexicali, Moussoul.

Llangollen: At any give time, a single employee was recording working times; the time series showed a drop in late 2017.

Lochemburg: Only one observation in this graph, with an excessively high value.

Mexicali: There were three employees prior to January 2019, with daily working times mostly under 300 minutes (5 hours); employee 8180 showed a peak in his time series in early 2019.

Moussoul: A single employee working in the mission, with a reported time series as a straight line; no reported data between October 2016 and January 2017.

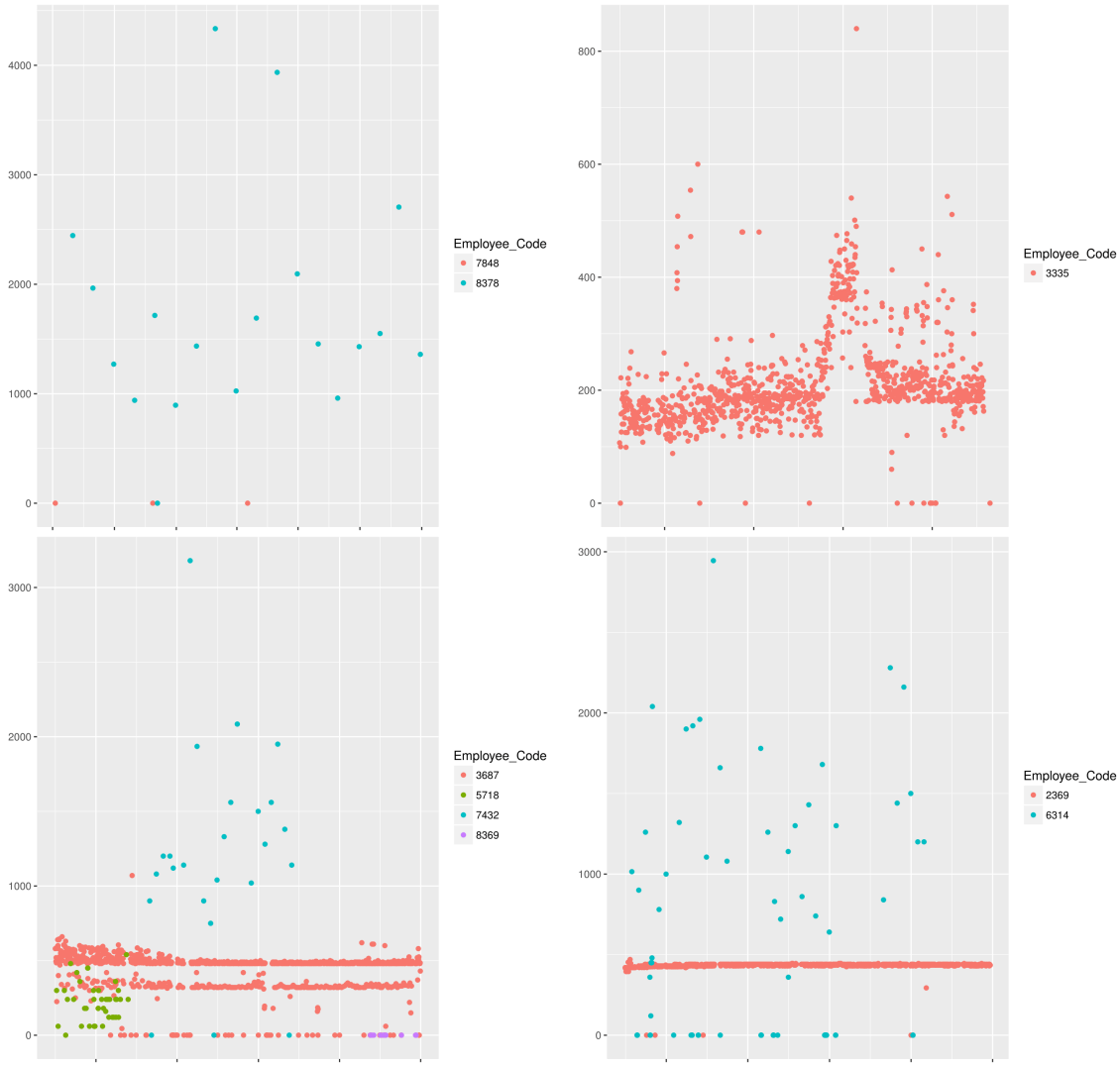


Figure 14: Daily working hours for 4 missions – VIII: Niroshimo, Nwakcot, Ouagasabre, Port Delgado.

Niroshimo: Employee 8378's time entries were very sparse and limited; employee 7848's time entries only had three records and all of them were 0s.

Nwakcot: A peak occurred in the time series of employee 3335 late in 2018, only lasting a few months; the overall trend of this employee's working times was going up slightly.

Ouagasabre: Employee 7432's time entries were very sparse and limited, but the reports include enormous (and impossible) amounts of minutes; employee 3687's time entries were fairly consistent and mostly oscillated between 0, 6 and 8 hours.

Port Delgado: Only employee 2369 reported constant working hours in this mission; employee 6314 had extreme working times that were usually longer than the other employee but he did not work regular days.

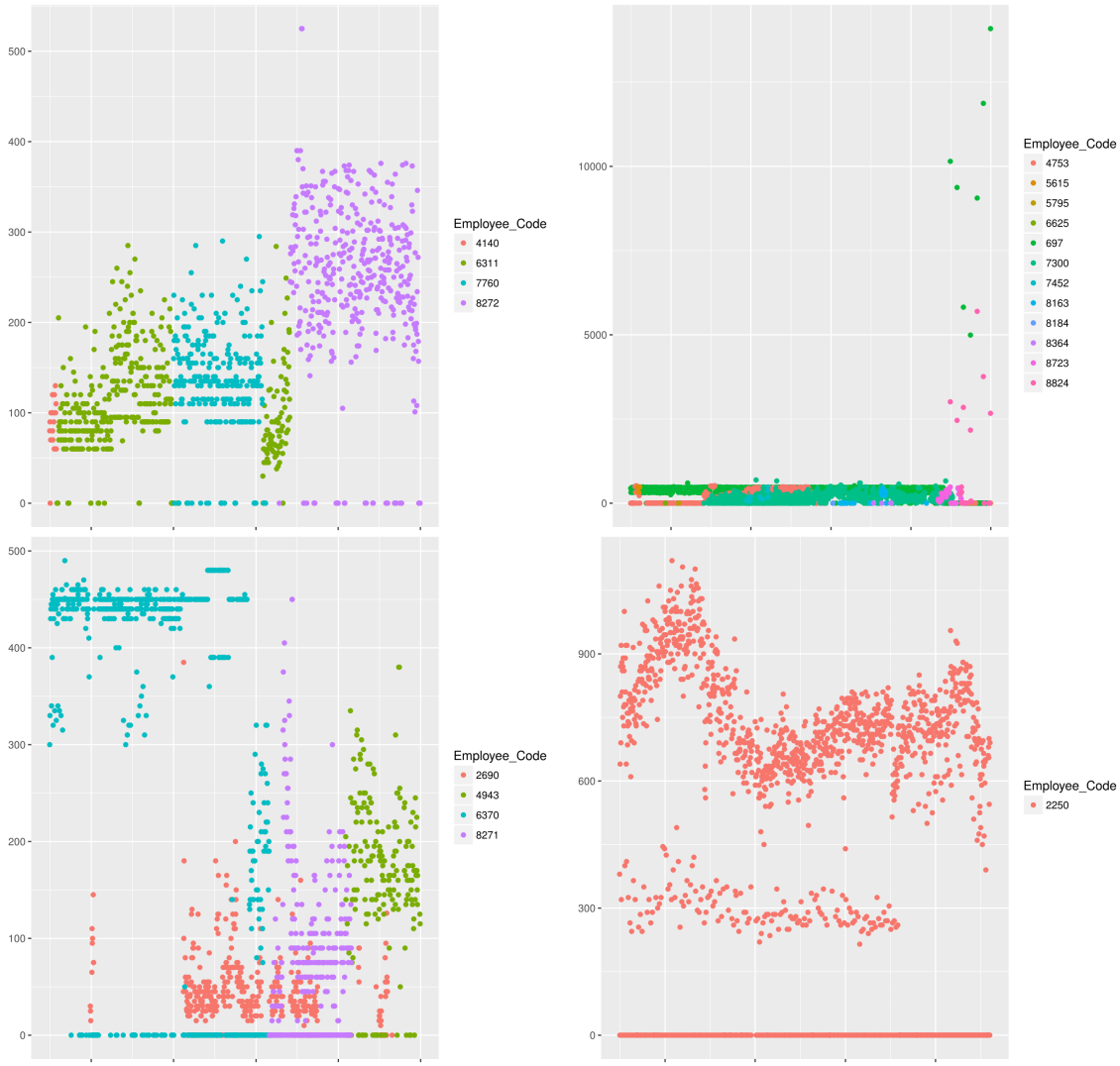


Figure 15: Daily working hours for 4 missions – IX: Poussanne, Ribeo, Riege, Sazal.

Poussanne: Four employees worked in the mission at separate times; the overall trend of the time series of all employees jumped up in early 2019.

Ribeo: All employees' time series were stable prior to July 2020; post July 2020, there were fewer entries and few employees working; extremely high values were recorded.

Riege: The overall working times have dropped since January 2019; the mission kept two employees working regularly between 2018 and 2020.

Sazal : This one defies any easy description.

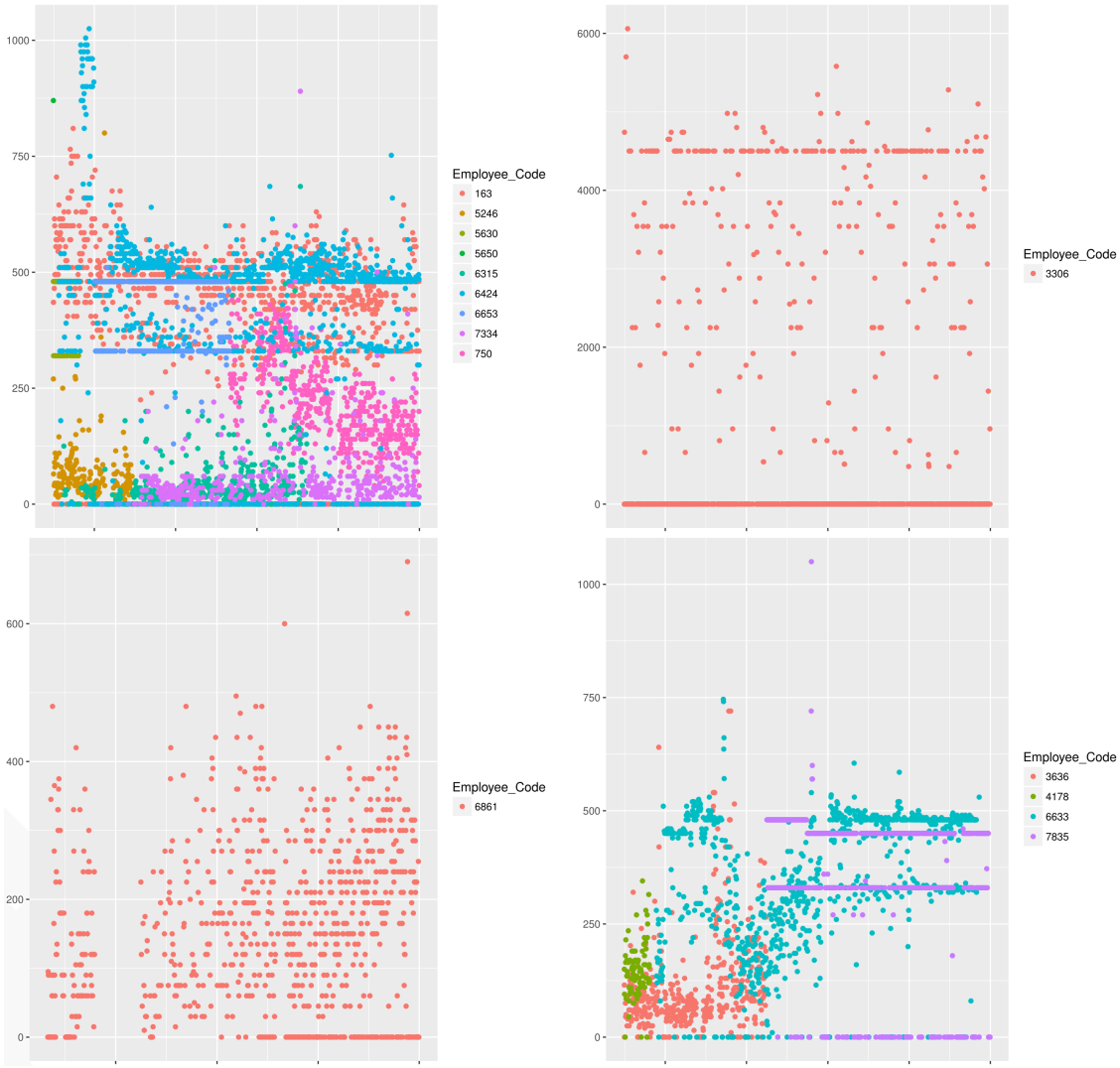


Figure 16: Daily working hours for 4 missions – X: Simhapura, Sofiya, St. Michelin, Teguz.

Simhapura: At least three employees were working in the mission at the same time; the time series of employee 9497 showed a decreasing trend after 2019, but not other employees.

Sofiya: A large number of observations were recorded as 0s; the employee regularly entered extreme values for his working time (above 1000 minutes per day).

St. Michelin: There was no clear pattern for this employee's working times; nothing was reported from October 2017 to March 2018.

Tegucigalpa: At least two employees were working in the mission at any given time; the time series for employee 6633 has a "V" shaped trend, with January 2017 as a turning point.

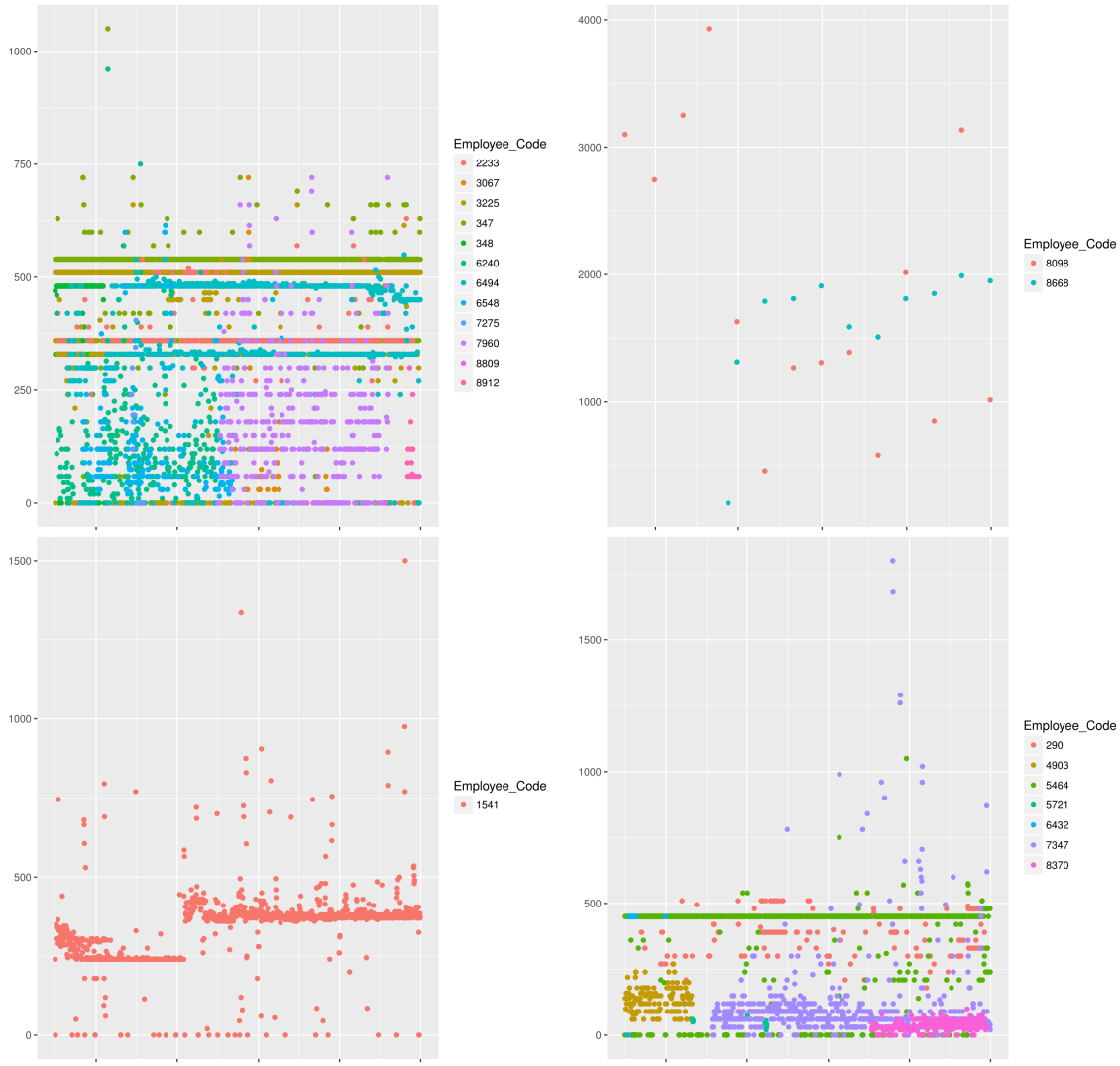


Figure 17: Daily working hours for 4 missions – XI: Tall Abib-Yafa, Thessalonia, Tia Juana, Wien.

Tall Abib-Yafa: Heaping is well-demonstrated by horizontal lines throughout.

Thessalonia: Only 2 employees, reporting at regular intervals, but with entries that are mostly incompatible with a daily format.

Tia Juana: The time series for employee 1541 is very clear: most of the entries were around 250 minutes (approximately 4 hours) prior to 2018, and around 400 minutes (approximately 7 hours) from early 2018 onwards; a large increase in working hours occurred early in 2018.

Wien: Three employees (4903, 7347, 8370) have reported hours which are fundamentally different from main employee 5464, and these "partial hours" are seemingly decreasing over time.

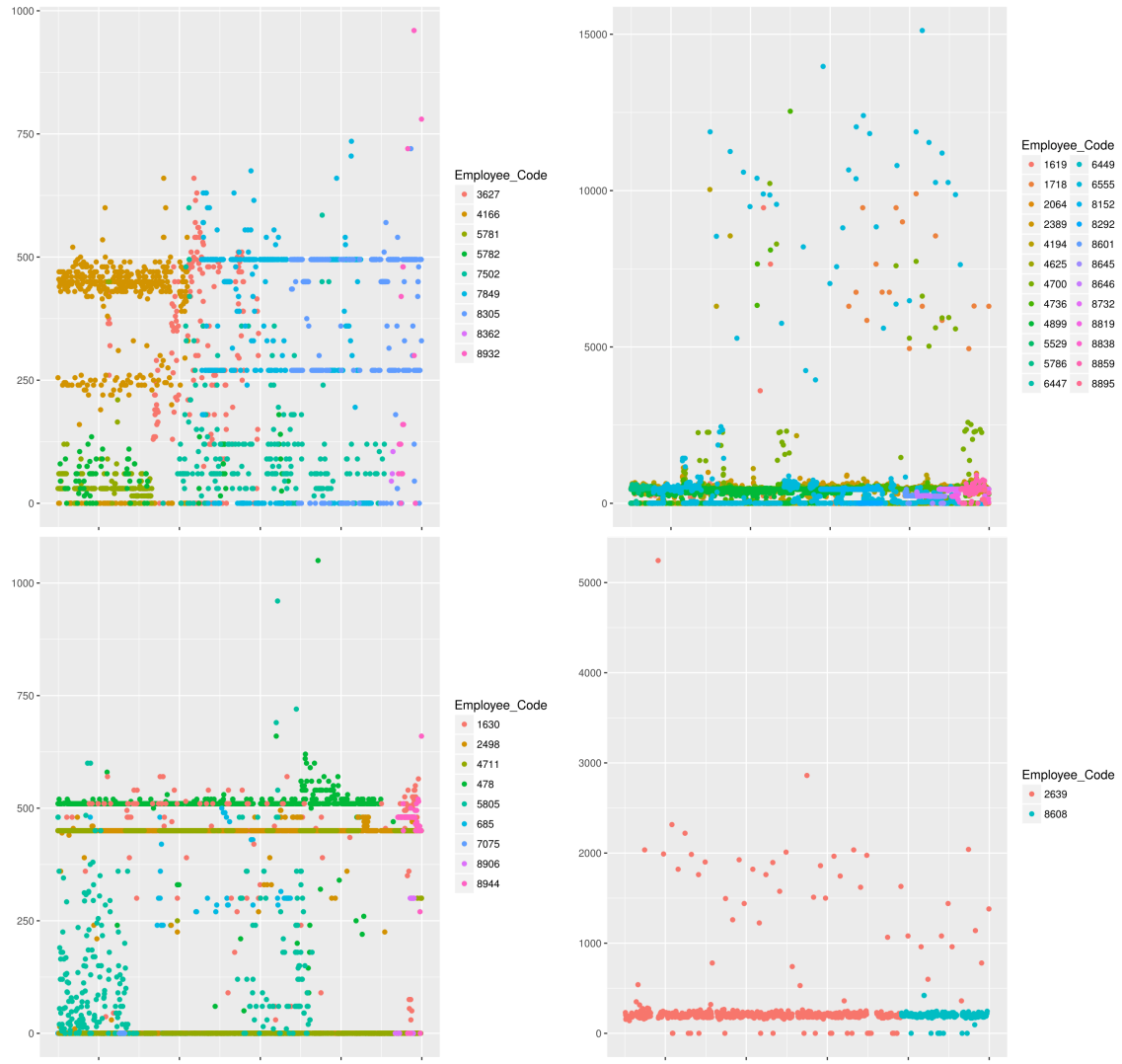


Figure 18: Daily working hours for 4 missions – XII: Davaka, Reme, Elgin, Tolosa.

Davaka: Employees 7649, 8305 and 4166 seem to be replacing one another, but their reporting styles vary.

Reme: A cloud of impossible results obscures patterns in smaller reported daily entries.

Elgin: The time series show many zero entries, and limit ranges at 480 and 510 minutes; employee 5805 did not report for most of a year, but when he did his reporting times were much more haphazard.

Tolosa: The two employees mostly reported at reasonable rates (around the 250 minutes a day range), but one employee regularly reports impossible days.

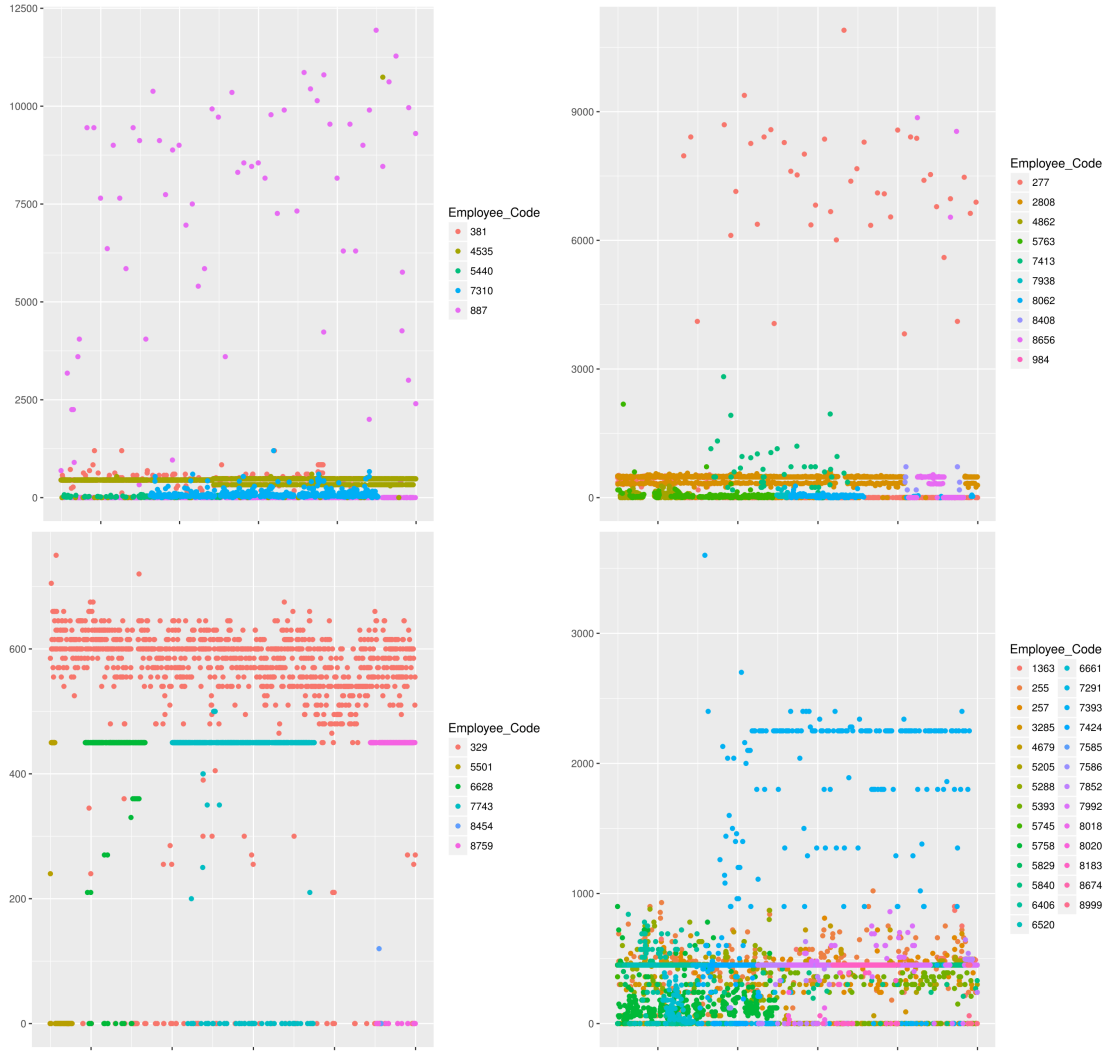


Figure 19: Daily working hours for 4 missions – XIII: Berograda, Bonaira, Caerdydd Newydd, Maximillian.

Berograda: Employee 887 had a large number of entries greater than 1440 mins, while employee 4535 and employee 381 mostly reported either 450 or 480 mins; the time series for the other two employees showed a smaller range, and was below 100 mins most of the time.

Bonaira: In mid 2017, employee 277 switched from $\sim$ 480 mins to either 0 or extreme numbers; only one employee (2808) reported regular working times (480 mins and 270 mins), and the other two employees had smaller rates (below 100 mins).

Caerdydd Newydd: At most two employees were only ever working at any given moment; on average, employee 329 worked around 600 mins daily and his time series was slightly decreasing; the other employee, whenever there was one, reported around 480 mins per day.

Maximillian: The time series of employee 7393 jumped at the end of 2017, from a smaller rate to above 2000 minutes, which lasted for three years until December 2020. The rest of the working team entered regular working hours (around 480 minutes).

Figure 20: Daily working hours for 4 missions – XIV: Partunis, Port of Wales, Praga, Tunis.

Paris: Almost all employees except for employee 4586 reported a similar range of times with very little variation, between 450 and 500 mins; employee 4586 only worked for about one year and his time series was around 250 mins.

Port of Spain: Two employees were reporting similar hours (either 270 or 500 mins); the other employees worked under 200 mins every day; employee 6253 showed a slight upward trend.

Praga: Unlike the majority of the mission staff, employee 6538 had extreme time values (either 0 or above 1000 mins) and stopped working in the mission in 2019; after his departure, only three employees reported times for the mission, two of which were reporting regular daily working hours (between 480 and 500 mins).

Tunis: Only one employee worked regular hours daily (between 300 and 500 mins); more employees were working from July to December on a yearly basis, which implies that the second half of each year is busier in general.

Figure 21: Daily working hours for 4 missions, without anomalous data – Adriata, Agruma, Damascia, Denniopolis. The removal of anomalous data makes it easier to spot patterns in the plausible data range.

3.4 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, and thus enter it incorrectly into the system (note that this misremembering may be due to a faulty memory, memory lacking in the necessary level of detail, memory influenced by biased perspectives, among other things);
- Employees are entering the data so that it reflects a certain desired reality, as opposed to what is actually occurring (they wish the mission to appear to be busy, for instance, and so they enter data that will reflect this wish);
- 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, a substantial sample of which are described below. Suggested tests to detect these, as well as

candidates found in the actual dataset, are provided whenever available. It should be noted that **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**, which is part of the problem: without external information and domain expertise, legitimate time series may look suspicious or anomalous.

For each of these scenarios, it becomes apparent that existing data is insufficient to distinguish between problematic and benign versions of the scenario. In principle, additional data could be gathered to improve the discriminating ability of metrics in these scenarios. However, even in this case, it would not be possible to validate the data in many of the described scenarios.

3.4.1 Scenarios Description

- Daily working hours on each task are entered by computing an average from the monthly estimated totals.

  Result: daily working hours should be evenly distributed for long stretches of time.

  Scenario 1: Alex forgets to enter his working hours. At the end of each month, the working time sheet has to be submitted, so he enters the same amount of time for each case/service he worked on.

  Scenario 2: Alex thought his work was identical to some extent so he entered the same working hours every day.

  Scenario 3: Alex did not spend much time on work but still entered 7.5hrs (for example) in the working time sheet.

  Challenges: no auxiliary (auditing) information exists to differentiate these scenarios from legitimate time series.

  Candidates: Addasibaba (employee 2692?), Atilanta (employee 2257?) (see Figure 7, p.16).

- Daily working hours on each task are entered using a reasonable guess based on past patterns and/or memory.

  Result: daily working hours are not likely to be evenly distributed.

  Scenario 4: Alan is an experienced employee, knowing how much time he usually spends on a regular case/service. He thus enters his approximate working hours differently depending on the case/service.

  Challenge: 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).

  Candidates: impossible to tell at this stage.

- Daily working hours on each task are entered with typos.

  Results: the time is entered in a different format (hours or days instead of minutes, say); outliers and anomalies in distributions, etc.

  Scenario 5: time should be a numerical value but it is entered as a character.

  Scenario 6: wrong position for the decimal. For example, 10.0 mins entered as 100 mins.

  Methods: check the consistency of the time formatting; generate the distributions of daily working hours for each mission, etc.

  Candidates: Kaartoom (employee 2174?), Limonia (employee 669?) (see Figure 12, p.21); Reme (multiple employees), Tolosa (employee 2639?) (see Figure 18, p.27).

- Daily working hours on each task are randomly entered, independently on the actual time spent on cases, services, and/or programs.

  Scenario 7: Adrian does not care about tracking hours and just makes up numbers for the time sheet.

  Methods: Benford's Law might flag some anomalies.

  Challenges: could be difficult to differentiate from an employee whose tasks provide for rather random times.

  Candidates: St. Michelin (employee 6861?) (see Figure 16, p.25).

- Daily working hours on each task are entered by copying another employee's time sheet.

  Result: daily working hours for two (or more employees) are identical for long stretches of time.

  Scenario 8: Amy doesn't know how to track hours so she copies her colleague's time sheet.

  Methods: comparisons of matching subsets of a mission's employee's daily reports.

  Challenges: it's quite conceivable that two employees have similar responsibilities, and that sequences of matching times are not indicative of invalid data.

  Candidates: Elgin (employees 478 and 1630?) (see Figure 18, p.27).

- 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.

  Scenario 9: April was too busy with her work and had no time to submit her time sheet.

  Scenario 10: Aimée worked at a mission with very little case/service/ program requests.

  Methods: look at distributions of employees' reported daily working times and seek distributions with large numbers of zero or near-zero values.

  Challenges: there could be many reasons why zeros could appear (both scenarios above, for instance, or if there is another dataset which records time spent on non-case/service/ program times, or as time off after overtime, perhaps).

  Candidates: Elgin (see Figure 18, p.27), as one of many.

- 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.

  Scenario 11: the weather forecast predicted that a tornado would hit Hokkaido, which led to a jump on the volume of passport related activities for the closest mission in that week. Ai decided that she would distribute her extra working time to the next week so that she could take a break at work.

  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.

  Candidates: impossible to tell at this stage without external or audit data.

- 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.

  Scenario 12: Ali does not know how to enter his working hours so he asked a colleague to submit the time for him.

  Challenges: without auditing data, this could look the same as a busy employee and a part-time employee.

  Candidates: impossible to tell at this stage without audit data.

- Daily working hours on each task are inflated before being entered.

  Results: an employee's records will be inflated.

  Scenario 13: Armstrong thinks he is underpaid and he enters more hours than what he worked.

  Scenario 14: Annick constantly mis-gauges how long it takes her to work on cases/services/programs and records higher values than the real values.

  Method: 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.

  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.

  Candidates: impossible to tell at this stage without external or audit data.

- 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.

  Scenario 15: Amber is an experienced employee and always works with 3 experienced co-workers (at her level). But the three quit at the same time and 2 new untrained employees are hired to replace the previous 3 positions. Amber has to spend more time to train the new employees, on top of doing her regular job.

Scenario 16: Amélie has been working at a mission for a few months and is getting better at her tasks, shaving about 20% off the time it used to take her to complete them.

Methods: identifying trend shifts using time series analysis.

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/decresase in reported times could also be associated with an increase/decrease in the number of specific cases/services/programs.

Candidates: Wien? (see Figure 17, p.26)

- The nature of the relationship between Borealia 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.

  Scenario 17: Ada is a full-time employee in Adriata. With more traffic and economic activities between Borealia and Côte Verte, a higher number of cases opened every day.

  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.

  Candidates: Sazal (see Figure 15, p.24), Adriata (see Figure 7, p.16).

- 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).

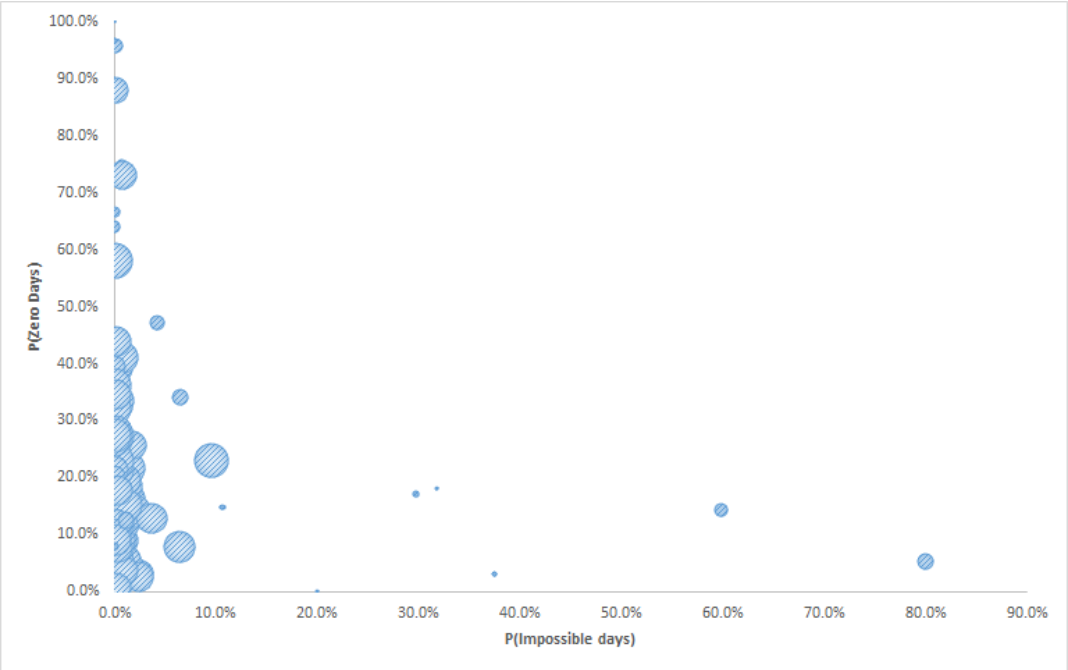
  Scenario 18: the number of employees in a mission decreased and the daily working times for each employee remained more or less the same.

  Scenario 20: the number of employees in a mission decreased and the daily working times for each employee decreased.

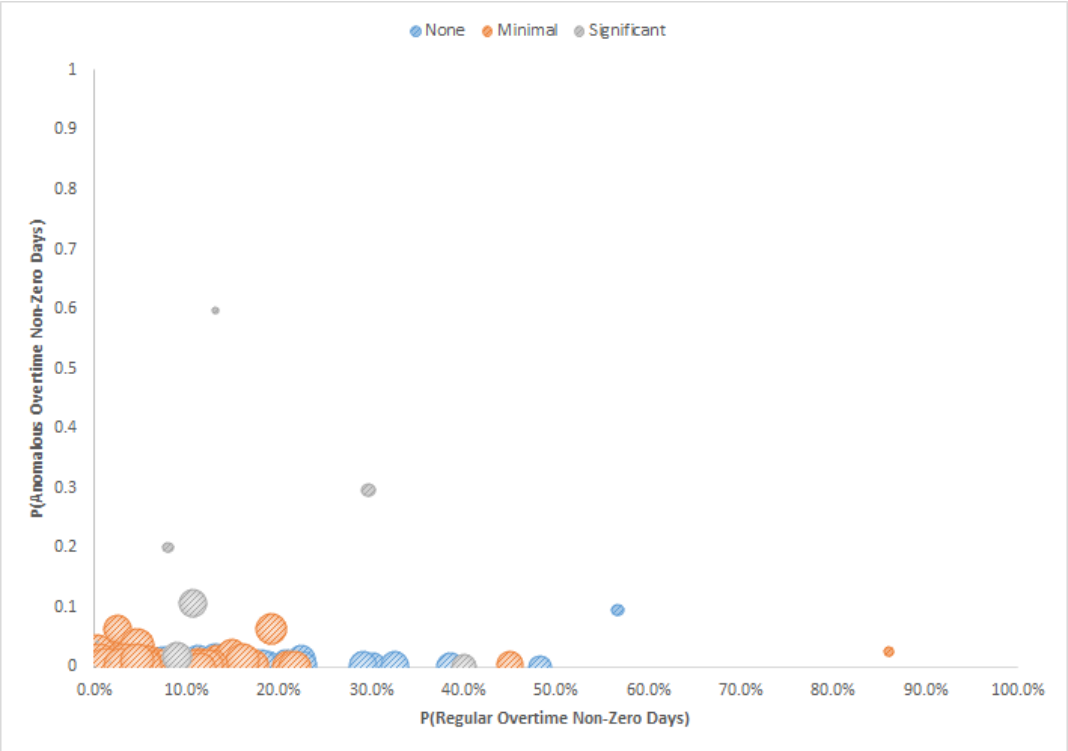
  Candidates: Agruma, Damascia, Denniopolis.

3.5 Plausibility of Work Hours

To analyze plausibility of hours entered, 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); a reasonable amount of time has been recorded (0—500 minutes); a plausible amount of overtime has been recorded (500—900 minutes), and an anomalous amount of overtime has been recorded (900—1440 minutes). Various bubble charts of the missions are shown in Figures 22 to 24. In Figure 22 (a), for instance, one would presume that in the absence of a valid explanation for zero days and impossible days, missions should find themselves in the lower left corner of the chart. Indeed, the further to the right we find a mission, the larger the proportion of entries recording days of more than 1440 minutes; similarly, the higher a mission



(a) Proportion of employee zero days vs. proportion of employee impossible days, by mission (the size of the bubbles is linked to number of entries).



(b) Proportion of employee anomalous overtime days vs. proportion of employee plausible overtime days, by mission (the size of the bubbles is linked to number of entries, the colour scheme is linked to the number of impossible days, as in Table 4, but Significant also includes Problematic missions).

Figure 22: Visualization of data validity at the mission level – I.

is located, the larger the proportion of entries recording days with 0 minutes. Missions which fall outside a reasonable range should be flagged for further analysis by domain matter experts.

In (b), we would also expect that missions should fall in the lower left corner: missions in blue are missions with no impossible entries (more than 1440 minutes); they seem to have a larger spread of plausible overtime entries than those missions for which at most 5% of the entries were impossible (in orange), which pile up closer to the origin, but with a tendency to have a higher proportion of anomalous overtime entries. As before, missions which fall outside a minimal threshold zone should be flagged for further analysis by domain matter experts (what these thresholds should be depend on the explanations for unlikely entries).

In Figure 23, the same data is presented separately, by proportion of impossible days (colour). The circles are still scaled by the number of entries at the mission level (at a relative scale for each graph), but they are not filled, making it easier to see the overlap. As the proportion of number of days with impossible entries (1440 mins and up) increases (going from blue to orange to grey to gold), we would expect the bubbles to rise (which corresponds to an increase in the number of anomalous overtime entries – 900 mins to 1440 mins). On average, this is indeed what we see, but without a good explanation as to why there are impossible days in the first place, this is only marginally useful in capturing invalid data.

In Figure 24, the proportion of entries with zero time is shown plotted against the proportion of entries with anomalous overtime entries (between 900 minutes and 1440 minutes), again separately by proportion of impossible days (colour). It is a different visualization of the same dataset for which a similar threshold argument holds; without external data and domain expertise, it is not possible to determine whether the data is valid at the mission level.

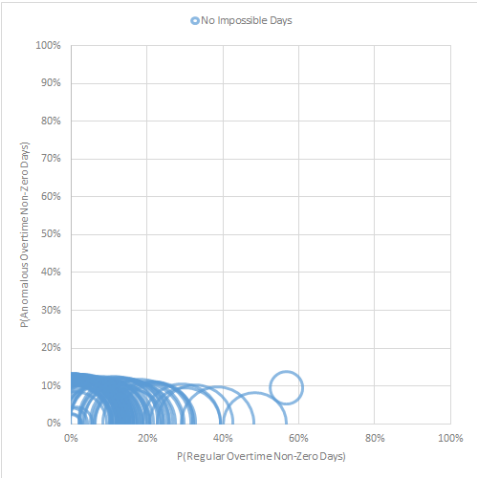
3.5.1 Data Validity at the Employee Level

Figures 25 and 26 repeat the visualizations of Figures 23 and 24, but at the employee level rather than at the mission level. Although there is more variability at the employee level (and it's easier to identify those employees who report impossible numbers on a regular basis), similar issues arise here – the lack of external data and (communicated) domain expertise make it difficult to flag employees for further study.

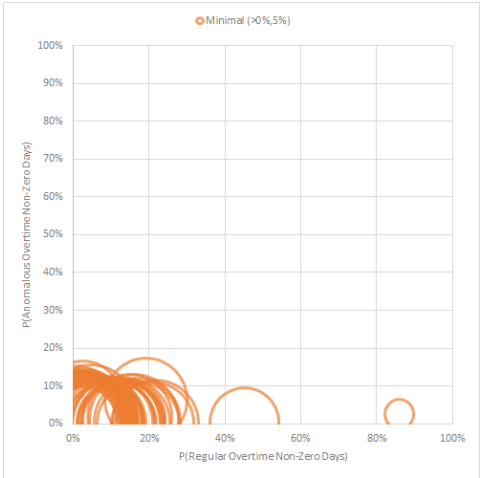
3.6 Recommendations for Improving Dataset Validity

Although it has already been noted that it is inherently difficult to validate mission data, it is still possible to improve on the current situation in this respect. More specifically, given the current issues described above, some measures can be undertaken to improve the overall validity over time, such as:

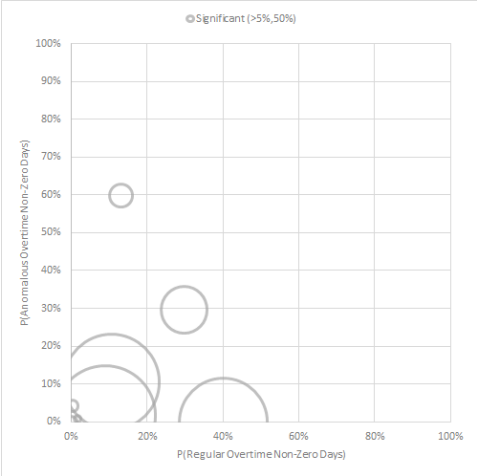
- including a list of mission-specific services: in our analysis, we have only looked at the combined time spent on cases, services, or program activities; if it is known that a given mission doesn't offer passport services, for instance, then this would provide analysts with an added basic check against invalid data by making sure no time is connected to passport services for that mission (continued on p.41);



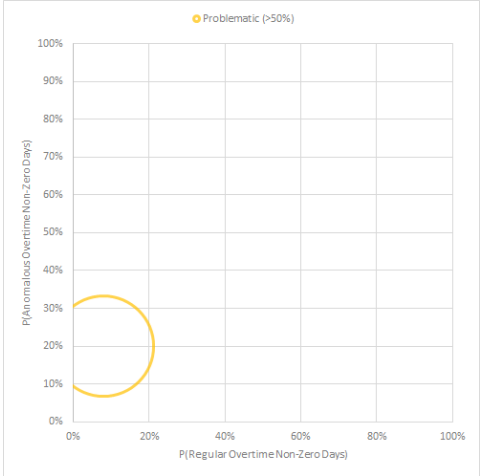
(a) No impossible days.



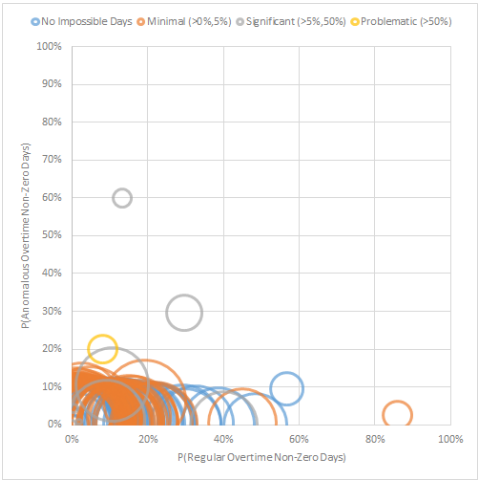
(b) Minimal (between 0% and 5%).



(c) Significant (between 5% and 50%).

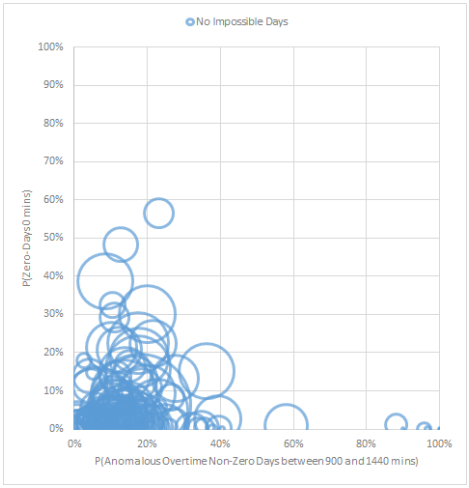


(d) Problematic (50% or more).

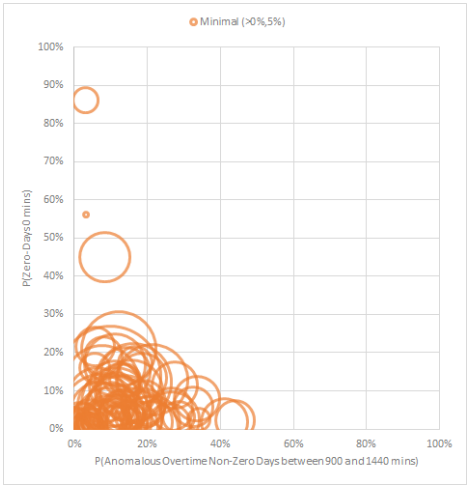


(e) Combined.

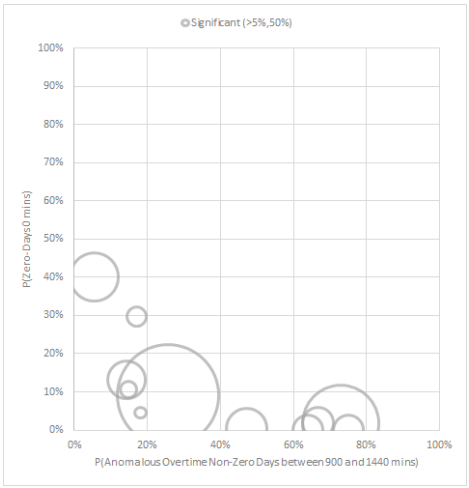
Figure 23: Visualization of data validity at the mission level – II. Proportion of anomalous employee overtime days vs. proportion of employee plausible overtime days, by mission and impossible days (the size of the bubbles is linked to number of entries).



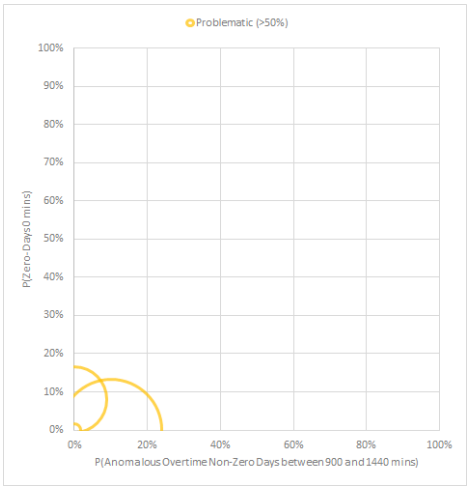
(a) No impossible days.



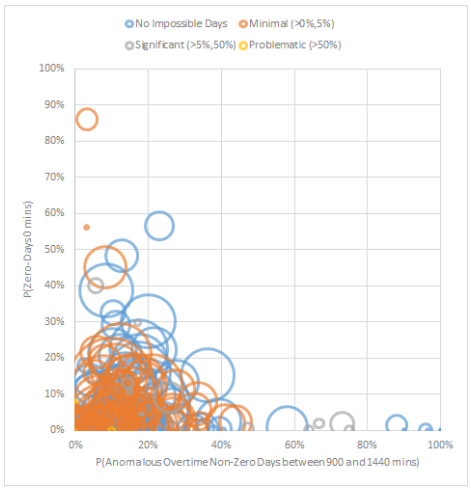
(b) Minimal (between 0% and 5%).



(c) Significant (between 5% to 50%).

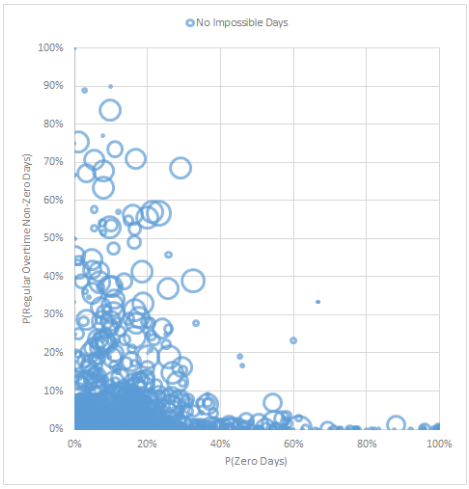


(d) Problematic (50% or more).

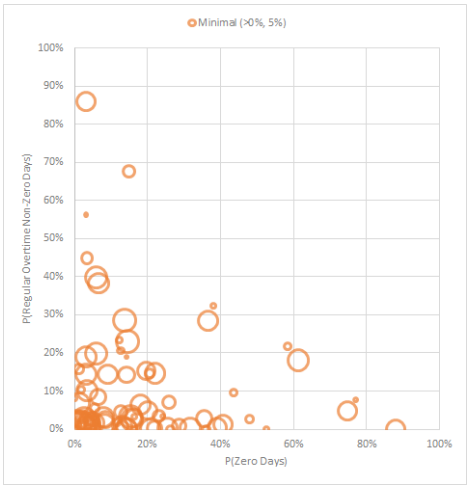


(e) Combined.

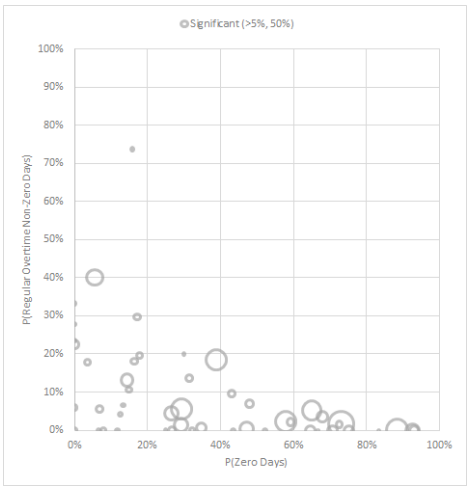
Figure 24: Visualization of data validity at the mission level – III. Proportion of employee zero-days vs proportion of anomalous employee overtime days, by mission and impossible days (the size of the bubbles is linked to number of entries per mission).



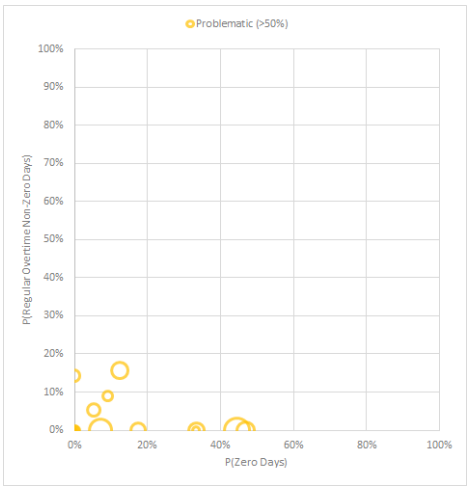
(a) No impossible days.



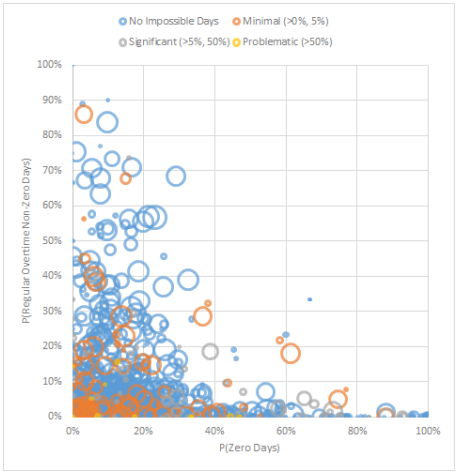
(b) Minimal (between 0% and 5%).



(c) Significant (5% to 50%).

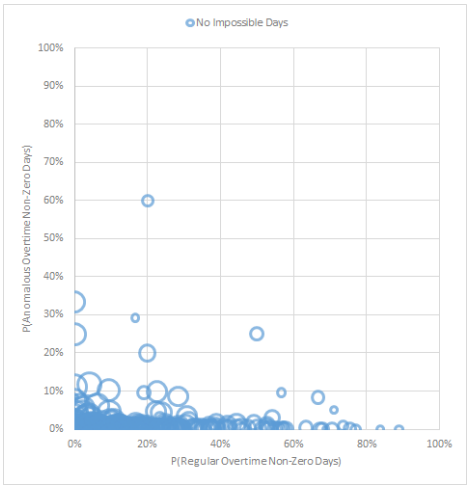


(d) Problematic (50% or more).

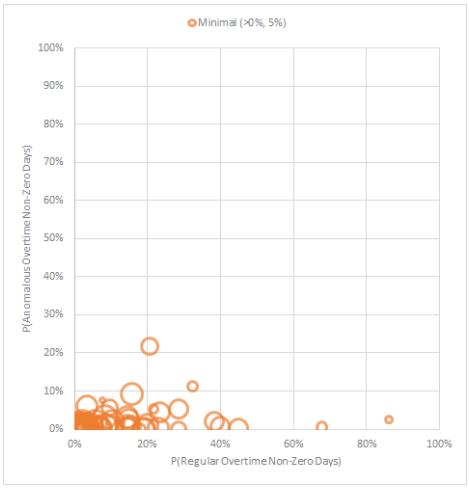


(e) Combined.

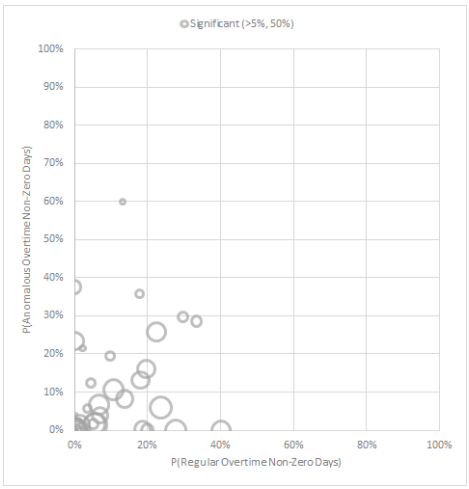
Figure 25: Visualization of data validity at the employee level – I. Proportion of anomalous overtime days vs. proportion of plausible overtime days, by employee and impossible days (the size of the bubbles is linked to number of entries).



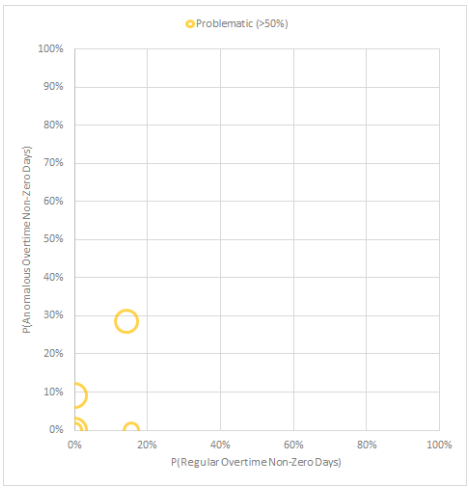
(a) No impossible days.



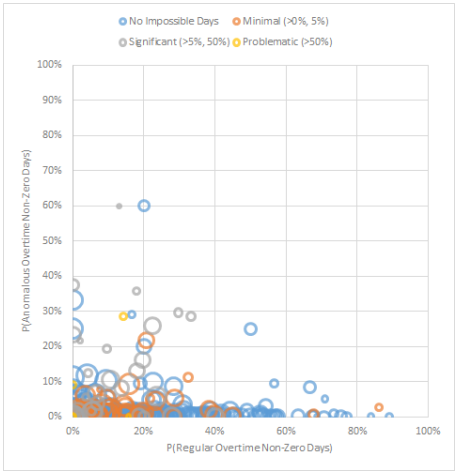
(b) Minimal (between 0% and 5%).



(c) Significant (5% to 50%).



(d) Problematic (50% or more).



(e) Combined.

Figure 26: Visualization of data validity at the employee level – II. Proportion of zero-days vs proportion of anomalous overtime days, by employee and impossible days (the size of the bubbles is linked to number of entries by employee).

- 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, or providing a uniform set of guidelines to follow: 0 entries should only ever appear if no time has been spent on a task and on a day for which it could, in theory, have been possible to spend time, and the overtime tracking system has to be overhauled so that employees do not have the possibility to report for the daily time worked by multiple people;
- tracking case work and service instances to specific employees as the work is being done: this would give the analyst the possibility of looking at mission – or employee – average time spent on case and service instances types and flag anomalous time reports for further examination;
- eliminating the need to report to the nearest minute: the data could support reporting to the nearest 5, 10, or 15 minutes, which could help eliminate some of the typos that were observed in the data.

4 Possible Metrics, Models and Analyses

Putting data validity issues aside for the time being, how can the type of mission data gathered to this point be used? Broadly speaking, CA wishes to use their data to make consulate management decisions relating to matters concerning:

- how BFO can provide an appropriate level of required services to Borealians given an available operating budget,
- which consulates should continue to exist (or where new consulates should be created), and
- the number of people required to properly conduct consular activities in each region.

Relating to these questions, high level metrics could in principle be created to assess:

- the overall effectiveness and efficiency of a mission,
- whether or not a mission has an appropriate number of resources, and
- whether or not a mission should continue to exist (and how other missions could be affected if it did not).

High-level metrics like these generally represent an aggregation of many underlying measures and types of analyses. Combining multiple measures is important, as a single type of measure taken alone frequently cannot distinguish between the multiple possibilities represented by this measure. For example, if a case takes a long time to complete, it might indicate that the case is very complicated and taking a long time to resolve, or that the case was given a low priority, or

that the person working on the case hasn't had much experience with this type of case. Thus, in constructing metrics we need to look at, and combine, multiple measures in order to get a clear picture of a particular situation, and also to rule out some explanations in favour of others. Once generated, metrics for a particular system can then be calculated over specific data in order to facilitate decision making under particular circumstances, primarily by acting as indicators that further investigation into a the situation could be beneficial.

We will consider possible strategies for generating each of the metrics described above, and assess the extent to which existing data can be used to generate these metrics. We will also discuss a strategy for generating a 'mission snap-shot' that reports both these metrics and a variety of basic analytics in order to generate an up-to-date picture of a given mission at a given moment in time. **It is important to note that, ultimately, CA must be the final arbitor of what should be used as the criteria for any given metric, given that they have the primary expertise on their consular system. However, for the purposes of illustrating possible metric calculations as well as the data required to support these calculations, some general considerations of possible criteria and calculations will be provided here.**

4.1 Effectiveness and Efficiency Metrics

Effectiveness is a measure of whether or not a particular goal has been achieved. Efficiency, on the other hand, is a measure directly related to the time and energy put into an activity.

In the case of effectiveness, the length of time required to achieve a particular goal is only relevant in so far as it is connected to the goal itself. For some goals, the time required may not be the primary measure of success. Similarly, the quality of the final result may also be more or less relevant, depending on the situation. To measure effectiveness, then, we must compare the results of the action with the specific goals connected to the action and determine how well these goals have been met.

In the case of efficiency, the situation is somewhat reversed. The activity itself must be defined, again, but the focus is not on the goal of the activity. Rather the principal concern is effort expended while the activity is ongoing. For efficiency, to generate a measure we must compare the time and energy involved in the activity with the time and energy required to carry out equivalent instances of the activity, either hypothetical or actual. The efficiency of the activty in question is then calculated relative to these other instances.

4.1.1 Effectiveness

Criteria If the overall goal of the consular network is primarily to serve and assist those who make use of its services (its clients), high level criteria for determining the effectiveness of the mission might include:

- How well the mission is meeting the needs of its clients (e.g. successfully providing the requested or required assistance)
- The level of satisfaction of the clients who are receiving this assistance

As has already been noted, effectiveness is about goal achievement and not time and energy put into an activity. However, in the case of mission assistance activities some consideration of time is

salient to effectiveness. Thus, one of the goals for a particular type of activity might be to complete activities of this type within a particular time frame (note that time related benchmarks used to calculate efficiency (see below) are not, in and of themselves, appropriate for use in calculating this aspect of effectiveness). However, relying on time related goals as the sole measure of effectiveness can introduce a number of problems related to a focus on activity completion time at the expense of quality of result. Thus, in addition to time related goals, each activity type will require the development of additional goals and associated measures.

An investigation into all of the appropriate goals and goal measures for each activity type is beyond the scope of this report. For the sake of discussion, three examples of possible effectiveness measures will be considered here: activity completed within acceptable timeframe, employee assessment of goal completion and client satisfaction rating.

Calculating Effectiveness For a given activity:

1. Obtain the completion time for the activity (see **Number of Complete Activities** for more details).

2. Obtain the completion time goal for this activity.

3. Calculate the timeliness score for the activity. For the sake of illustration, one possible strategy for this is as follows:

   - IF activity completed before acceptable time frame has elapsed THEN timeliness $= 1$
   - IF activity completed within acceptable time frame THEN timeliness $= 0$
   - IF activity completed after acceptable time frame has elapsed THEN timeliness $= -1$

4. Obtain the employee rating of goal, completion – e.g. rating on scale from $0 - x$.

5. Obtain the client satisfaction rating – e.g. a client assessment of satisfaction on a rating scale from $0 - x$.

6. Depending on which of these measures are considered more important to effectiveness of the activity, set weights that will emphasize one or the other.

7. Normalize measures and add together each weighted measure to get an effectiveness score for that activity.

8. Average the effectiveness scores at the desired level of granularity to get an effectiveness score for the mission for a particular time range.

Although this example shows a calculation at a high level of granularity (each activity), it's important to note that it's possible to measure benchmarks, employee assessment measures and client satisfaction at different levels of granularity. For example, the benchmark measure could be set at a monthly level (e.g. this many opened-and-closed cases over the month), employees could provide a goal completion assessment once per month and measures of client satisfaction could be collected not for every case but rather in a monthly sample of client satisfaction.

Required and Available Data Determining effectiveness requires measures relating to goal achievement, some of which have been illustrated in the provided metric calculation. At this time, none of the data required to calculate the effectiveness metric described above is fully available. Specifically, time to complete specific activity types is not available, a count of the number of cases opened-and-closed is not available, and there are currently no consistent and methodically collected measures of goal achievement and client satisfaction over particular time ranges. With respect to time to complete an activity, in the case of effectiveness this will likely not be 'number of hours' but, more relevantly for effectiveness, the length of time required to compete the activity (e.g. a case might require a total of 5 hours to complete, but these 5 hours could be spread out over 15 days – it is likely that the 15 day measure will be more salient for goals relating to case completion times).

4.1.2 Efficiency

Criteria Criteria for determining the efficiency of a mission are most logically based on measures of throughput- e.g. number of cases closed, number of clients assisted over a particular amount of time. Thus, the primary challenge with efficiency measures is not determining what to measure, but determining what is appropriate for comparison when assessing these throughput measures. In the case of the consular network, in particular, we know that what should be considered high throughput is highly mission dependent, because a situation that could be easy to resolve in one location might be very hard to resolve in another. Even within the same mission, assistance of the same type might be easier or harder to provide depending on the circumstances involved.

There are two, non-exclusive options for taking this into account:

- Take the difficulty level of the assistance into account directly in some way, e.g. by having mission personnel rate the difficulty of each type, or even instance, of assistance provided at their particular mission – for example, by creating a mission specific rating scale.
- Take care to make fair comparisons by comparing only genuinely similar instances. For example, categories of missions which can be fairly compared with each other should first be generated, and then comparisons only made within these missions categories. This may potentially be achieved by defining a number of relevant properties of missions, and then clustering missions using these properties in order to determine which should be placed in similar groups. See Section 4.1.3 for a further discussion of possible and appropriate clustering strategies.

Related to this last point, when calculating an efficiency metric it is also important not to merge dissimilar activities together – e.g. adding the times of disimilar activities and taking the average – since such a grouping would not make it possible to compare activities only with other similar activities. Specifically, in the case of the consular network, based on the information given to us, because different types of cases and services are very different, when looking at time required to carry out these cases and services the different types of cases and services should be considered separately. Also because of this, assuming employees work on different types of cases and services, it doesn't make sense to look at how efficient a particular employee is overall, because in this case we would again be forced to merge categories. This strategy is further problematic because of scenarios where one employee consistently works on a type of case that takes longer to resolve

than the type of cases another employees work on – in such scenarios it would falsely appear that one employee was less efficient than others, simply because a comparison was not being made to an appropriately similar situation.

Calculating Efficiency

Number of Complete Activities: First, in what will be a basic component of several metrics, consider strategies used to count activities like cases (and any activities like cases that are distributed across multiple days – this approach may not be required for services, if they are by definition opened and closed in one session. Program activities can be counted in this way if they can be viewed as a collection of discrete activities).

1. For a given range of time, separate the activities in question (cases are here used to illustrate the approach) into:
    - The number of open-and-closed cases (i.e. cases that are both started and completed in the range of time), and
    - the number of ongoing cases (i.e. cases that were opened before the start of the time range, closed after the end of the time range, or both).
2. Calculate the percentage of open-and-closed cases in the defined range, relative to all cases. Note that if the percentage is below some pre-determined threshold then this data range should not be used to calculate benchmarks or other measures.
3. Assuming the data range provides an adequate sample of open-and-closed cases, divide these into case types, and for each open-and-closed case in each category, calculate the time required to complete that case.

Activity Category Time Measures: For the mission and time range of interest then calculate, for each activity category, the average time required for that activity over the time range.

Activity Time – Internal Benchmark: A simple internal benchmark for a particular activity category will be the average time taken for activities of that type over a selected range of time within the mission – here, a wide time range should be selected, and consideration should also be given to selecting a representative period of time. More sophisticated measurement strategies may be used, but this will be sufficient for current considerations.

Activity Time – External Benchmark: Similarly, an external benchmark for a particular activity category would be the average time required for that activity over a selected time range for missions that have been included in the same cluster as the current mission (See Section 4.1.3 for further details on constructing clusters).

Compare and Sum: For each activity category, determine the difference between the activity category time measure and the selected benchmark time. Normalize the value by dividing by the benchmark time. Sum the resulting values. Divide by the number of activity categories.

Note that, whether or not an internal or external benchmark was used, when comparing efficiency scores, missions should only be compared to other missions in the same mission cluster.

Required and Available Data The primary data required for this metric is the amount of time per (completed) activity of a particular category, over a particular defined data range. This may be of varying levels of granularity. In contrast to the effectiveness metric, the key measure for activity time here is 'number of employee hours invested in the activity' rather than 'length of time required to compete the activity' although this second measure may also be factored into the final metric. As it stands, this data is not currently available. Data is available for total amount of time worked over a given data range (e.g. a month), but this is not divided into separate case and service categories. As well, time to complete cases is not currently available, as only the start date of cases is recorded.

Using external benchmarks and comparing across missions further requires the creation of mission clusters. For more information on the data required to carry out clustering, and its current availability, see the following section, 4.1.3

4.1.3 Mission Clustering

Determining efficiency requires a comparison between missions. However, the high variability of circumstances across missions, as highlighted by the time series provided in Section 3, makes it difficult to select missions that can legitimately be used for comparison purposes. One strategy that can be used to address this issue is a data mining technique called clustering.

In clustering, relevant properties are used to create novel sub-groupings among a collection of objects – in this case missions – such that those within the sub-grouping are more similar to each other than those in other sub-groupings. In the case of missions, assigning relevant properties (both intrinsic and extrinsic) to missions and then clustering over them can result in suggestions about which missions can legitimately be compared with each other.

Importantly, here, clustering is an unsupervised machine learning technique, which means that it will not make assumptions about which properties of a mission make it more or less similar to other missions. Rather, it will place missions into groups based on overall similarity between miisions, instead of focusing on only one attribute of the missions. Rather than, for example, a priori assuming that missions that are geographically proximate are similar to each other, clustering will group the missions instead by considering all of the multiple factors used to describe the missions, calculating mission similarity based on all of these factors. In doing so, the clustering algorithm may happen to group some missions together that are geographically proximate, but may also group missions that are geographically far from each other into the same cluster, because other factors (e.g. the size of the mission and the type of government in power in the host country) are more salient in determining similarity.

This example underscores another strength of clustering, which is that by using this technique it is possible to take into consideration a wide variety of attributes connected in some way with missions (e.g. a variety of aspects of the country in which the mission is situated), rather than

focusing on only one or two properties. The clustering algorithm will attempt to take into account all of the included factors.

Based on discussions with CA, the following properties of missions and countries might be useful in constructing mission clusters for the purposes of mission comparison:

- Mission properties:
  - Mission hardship rating
  - Number of employees
  - Employee experience measures
- Country properties:
  - Corruption index
  - type of bureaucracy
  - Relationship with Borealia
  - Security level
  - Environmental conditions
  - Human rights record
  - Languages typically spoken
  - Telecoms level
  - Cultural aspects
  - Type of bureaucracy
  - Technology literacy index
  - Size of country
  - Service availability
  - Distance from Borealia
  - Emergency proneness

In order to create clusters using these attributes, an appropriate feature based clustering algorithm must first be chosen, and missing values imputed. Generally speaking, some trial and error is then required to select the best clustering strategy and algorithm parameters, and multiple clustering runs generated to consolidate clustering results. Ultimately, the outcome will be a division of the full mission complement into sub-groups of missions that are more similar to each other than the missions placed in other sub-groups, with respect to the properties selected for consideration by the algorithm.

In terms of feasibility of carrying out clustering with currently existing data, discussions with CA suggest that many of the measures listed above could potentially be collected for each mission or country, which further suggests that some mission clustering could be carried out once this data was collected. The appropriate application for these clusters will depend on the particular mission features selected, and is something to be further discussed with CA, prior to proceeding with this analysis.

4.2 Resource Sufficiency Metric

Drawing on the calculations of effectiveness and efficiency metrics, we can then ask: How might it be possible to determine a mission's level of resource sufficiency? In other words, how can we determine if a mission has insufficient resources to properly do its job? Or too many resources? Or exactly the right number? Answering this question is not always straightforward or obvious – for example, it's possible for a mission to be highly effective (in the sense of achieving mission goals and successfully assisting clients) while having too many resources and, in the reverse, for a mission to have not enough resources and still be effective – albeit at a cost, and possibly only over the short term.

4.2.1 Criteria

Some possible indicators that a mission has insufficient resources are:

- amount of overtime in the system,
- the number of 'backlogged', delayed or long running cases,
- the amount of time it takes for clients either to initially receive assistance, or recieve ongoing assistance (i.e. wait time), or
- a low effectiveness rating.

It is important to note, however, that although all of these are possible indicators of insufficient resources, none of them are definitive. They do not necessarily distinguish, for example, between a situation where, for whatever reason, a different person could have done the same job more quickly (suggesting that the problem is not too few resources but a need for different resources) and a situation where only adding more people will make a real difference to the situation.

Thus, in addition to the above criteria, the **efficiency** of the mission must also be taken into account. If all of the above indicators are in problematic ranges, but the efficiency of the mission is low, this suggests that the solution is not to increase resources, but to increase efficiency. Based on this, the main criteria for a resource sufficiency metric are a combination of mission effectiveness, ratio of hours worked to number of employees, and mission efficiency.

In connection with this, consider the following scenarios:

Overabundance of Resources In general, a very simple indicator of over abundance of resources would be if the ratio of the total number of hours worked divided by the total number of people working is relatively and consistently low (e.g. lower than the number set for typical expected work hours per full time equivalent positions).

Note that at this point it might be tempting to remove resources until this ratio is at the desired level. One challenge with this, however, is that the performance of the mission may be a reflection of a number of interdependent factors – e.g. of how people are working together – and so it may not be the case that simply removing people will result in a mission that functions equally well while also moving into the 'sufficient resources' category. In this case, a distinction should be made between a mission that has an overabundance of resources and is highly effective vs one

with both an overabundance of resources and low effectiveness. The risk of disrupting an already low-effectiveness mission would presumably be less of an issue than disrupting a highly effective mission.

Sufficient Resources The main indicator of a mission having sufficient resources is if it has a high effectiveness metric, and a relatively constantly high ratio of work to people. In such a case it is still possible that the mission might have a low efficiency metric, which means that it might theoretically be possible to increase efficiency. In this case, increasing efficiency could have the effect of shifting the mission category into one where it has an overabundance of resources. However, as noted above, it should not be assumed that resorting to simple strategies such as removing people to change the ratio of work hours to people will necessarily have this effect.

4.2.2 Calculating Resource Sufficiency

1. Using number of people who have worked over a certain amount of time and number of hours worked, calculate the average number of hours worked per employee. Based on a chosen threshold, turn this into a categorical value (e.g. under a certain theshold = low hours/employee, within the threshold = expected hours/employee, above the threshold = high hours/employee).
2. For the same time range, determine the effectiveness metric and the efficiency metric for that mission. Determine acceptable thresholds for each of these metrics (e.g. above 0.5 is effective or efficient, below 0.5 is not, or choose more fine-grained divisions).
3. The number of possible categories will be: hour ratio categories × effectiveness categories × efficiency categories. However, these may then be grouped into larger relevant categories.
4. Flag particular categories as requiring further investigation.

Some categories of particular interest might be:

- High work-hours ratio:
  - + typical/high efficiency, typical/high effectiveness → More resources required
  - + low efficiency, low effectiveness → Non-resource issues
- Typical/Low work-hours ratio:
  - + high effectiveness + high efficiency → Suggests appropriate resource level
- Typical/Low work-hours ratio:
  - + high efficiency + low effectiveness → may be trading off effectiveness for efficiency

4.2.3 Required and Available Data

As two of three principal components of this metric are efficiency and effectiveness, see the discussion of those metrics for data required to calculate them. In addition to this, hour ratio measures are required. This hour ratio data is currently available.

4.3 Existence Metric

While an effectiveness metric might indicate areas for improvement in the operation of a particular mission, or conversely, missions with strategies that could be shared with other missions, this does not address issues relating to whether or not a particular mission should continue to exist in a particular area (or conversely, whether or not a particular area requires a new mission). In general, current or past performance metrics are likely not good indicators of which missions should stay open, or be closed, because each mission does not operate in a vacuum. Thus, it's not clear that, for example, closing one mission and giving its resources to another will really result in a more effective system over all, as the effects of closing a particular mission may have unanticipated side effects.

Ideally, we would like to know what effect closing a mission or opening a new mission would have on the network of missions. For example, we might consider: if we close this mission, how far would someone have to travel to get to a mission which offers comparable services? Thus, in the context of the network of missions, we might say that a mission should continue to exist if its absence would cause a significant problem for people who would potentially need to use the mission or who work at other missions.

4.3.1 Criteria

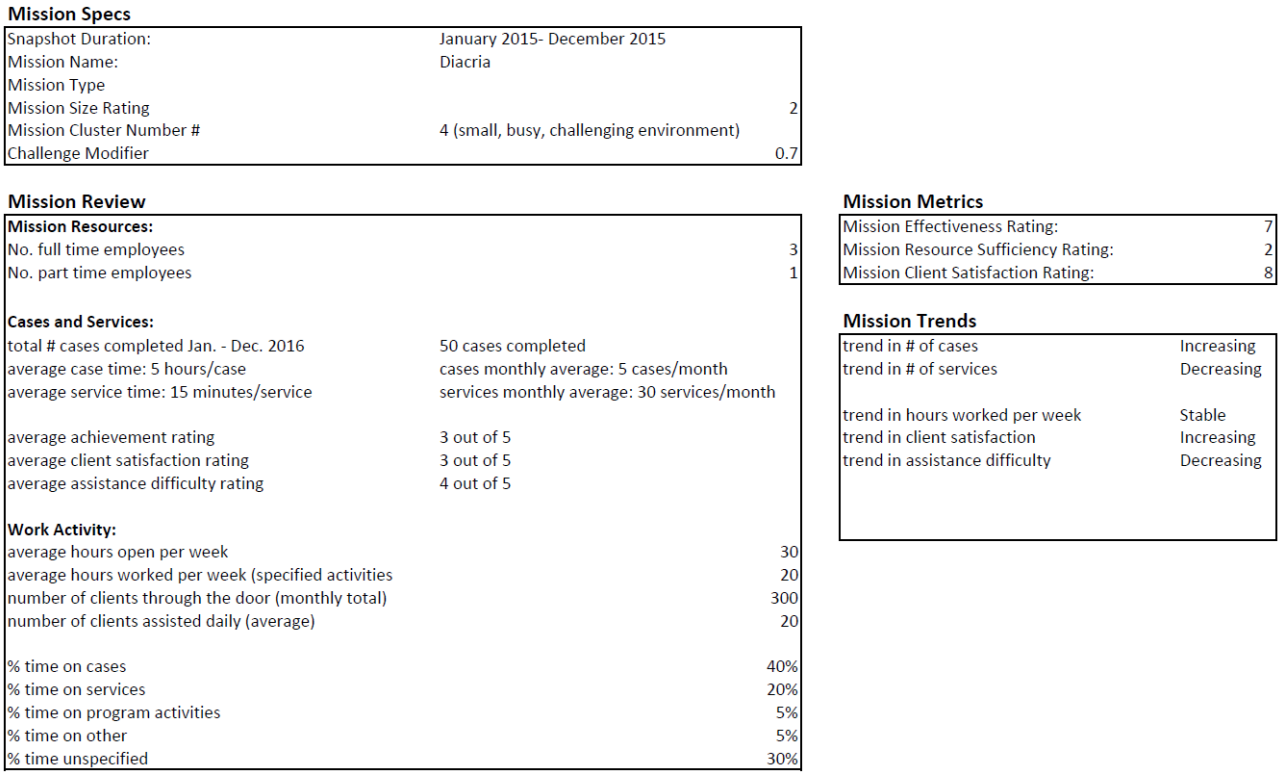
To answer questions like 'How far would a person have to travel (or how much effort would they need to put in) to get access to the following service?' and 'How much additional traffic would surrounding missions receive if this mission were to be closed?' we would need information about the likely number of services that a current mission is performing, as well as how busy it and the surrounding missions are. We would then project the likely increase in number of services under scnearios of interest using various heuristics (e.g. if a person is in location A, they will most likely go to the closest mission that provides a particular service).

To measure this we need to define what would constitute a significant problem for people and mission workers, and then what we would need to measure this. Relevant criteria might include:

- How busy is the mission currently?
- How busy is the mission likely to continue to be?
- Are there other missions that people could use instead?
- Is the mission providing critical or unique services that cannot be supplied elsewhere, or by someone else?

4.3.2 Calculating Existence Metric

We can address questions like the ones raised above by construct a graph network based on distance (or travel time, or travel effort) between missions. The results of this might then be used to construct an existence metric. The details of such a graph model are beyond the scope of this report. However, the general strategy is to construct a graph of missions, with missions as the nodes and the edges representing some kind of relationship between the missions – e.g. distance between missions (such a graph can be represented as a matrix). Once constructed the removal



Mission Specs

Snapshot Duration:	January 2015- December 2015	
Mission Name:	Diacria	
Mission Type		
Mission Size Rating		2
Mission Cluster Number #	4 (small, busy, challenging environment)	
Challenge Modifier		0.7

Mission Review

Mission Resources:		
No. full time employees		3
No. part time employees		1
Cases and Services:		
total # cases completed Jan. - Dec. 2016	50 cases completed	
average case time: 5 hours/case	cases monthly average: 5 cases/month	
average service time: 15 minutes/service	services monthly average: 30 services/month	
average achievement rating	3 out of 5	
average client satisfaction rating	3 out of 5	
average assistance difficulty rating	4 out of 5	
Work Activity:		
average hours open per week		30
average hours worked per week (specified activities		20
number of clients through the door (monthly total)		300
number of clients assisted daily (average)		20
% time on cases		40%
% time on services		20%
% time on program activities		5%
% time on other		5%
% time unspecified		30%

Mission Metrics

Mission Effectiveness Rating:	7
Mission Resource Sufficiency Rating:	2
Mission Client Satisfaction Rating:	8

Mission Trends

trend in # of cases	Increasing
trend in # of services	Decreasing
trend in hours worked per week	Stable
trend in client satisfaction	Increasing
trend in assistance difficulty	Decreasing

Figure 27: Sample dashboard for a fictitious mission.

of a particular mission in this graph would be defined to have particular effects on the connected nodes, which might then have ripple effects on the nodes connected to these nodes. In this way, the effect of removing missions on other missions can be calculated based on the resulting properties of the missions within the graph model.

4.3.3 Required and Available Data

It would be possible to incorporate a diverse range of data into the graph model, but a minimum requirement would be data relating to the relationship between missions in the network (e.g. a distance relationship) and data on an aspect of the missions that would likely be affected by the addition or removal of missions (e.g. number of clients and types of service provided).

4.4 Mission Snap Shot

Due to challenges relating both to confirming the validity of the 2016-2020 data, and, more critically, the absence of data required to calculate the provided metrics, the proposed metrics were not calculated for the current data. However, assuming the required types of data are collected at some point, it would be possible to construct a dashboard which could provide a snapshot of each mission by calculating measures, metrics and creating data visualizations over a specified range of data. A sample dashboard sketch is included in Figure 27.

4.5 Recommendations for New Types of Data to be Collected

In order for the metrics and dashboard discussed above to be created, a number of types of data are required that are not currently available in the PIMENTO database alone. Discussions with CA indicate that some of these types of data are possibly available from other sources. However, much of the required data will need to be collected on a moving-forward basis. There are two main categories of data to be considered in this respect, client experience data and work activity data. Specific details of the data required have been provided for each metric, but will be summarized here.

Client Experience Data There is currently relatively little structured data available on client experience at missions, although there is some general questionnaire data available which suggests that clients are broadly satisfied with the assistance they receive from missions. There is also relatively little structured data reflecting how successful a mission has been in providing the required type of assistance, although existing case notes may provide some information on this topic, and it may be argued that, in some situations, simply closing the case or concluding the service transaction indicates that the required assistance has been provided. It should also be noted that it is possible to successfully provide assistance even if client satisfaction is low (e.g. a new passport has successfully been issued even if the service interaction itself was disagreeable to the client). In the absence of this type of data, it is tempting to simply use throughput measures to determine effectiveness. However, this is conflating effectiveness with efficiency.

Work Activity Data As has been stated, the primary work activity data that needs to be collected in order to make the calculation of these metrics possible is the start and stop time of activities (cases, services, program activities), and, by extension, the time associated with each instance of these activities. As well, it is important to collect employee time separately from time worked for the mission by local workers who have no employee number. This is not to say that these hours should not be accounted for, but they should be entered into the system separately, under a separate code. Employee experience data or knowledge (e.g. measures of case difficulty or approriate case goals) should also be collected.

5 Conclusion

The variability of activities from mission to mission, and even, within missions over time, is a fundamental quality of mission work, and one which makes data validation extremely challenging. A major goal of the first section of this report was to provide an extensive explanation, with detailed examples, to demonstrate convincingly the reality of this challenge. However, a major take home point from this exercise must also be that internal data about mission activities can be supplemented with external data and knowledge of missions in order to provide cues to interpret what would otherwise be ambiguous data.

As discussed in the second half of the report, using the currently gathered data for decision support is also hampered by data collection decisions made when the intended functionality of the system was specifically to support working with and assisting clients. This focus meant that some

types of data that are basic to calculating performance metrics (e.g. throughput) have not been collected. However, as has hopefully been demonstrated, if some changes can be made to the way data is collected in the system, and some new types of data collection added, metrics can be calculated that will greatly enhance decision making capabilities relating to consular management.

Thus, as noted at the beginning of this report, if systems are, first, redesigned to fully capture the required data and then combined with existing external data, there is the potential to calculate a number of potentially useful mission-level metrics, as well as, more generally, create a mission snap-shot that can provide an at-a-glance summary of relevant mission level information. Both of these system outputs could then be used to facilitate decision making relating to mission management.

A Results of Basic Data Checks

A.1 Basic Data Assessment Results

No incorrect data types or ranges: The data entered into fields was consistent with the designated data types and ranges for each field. For example, only numeric characters were entered into the fields representing time spent on cases or services and number of cases opened, and no negative numbers were entered in these fields. Similarly, only character values were entered into fields providing categorical information.

Missing values: Some case and service category related field cells had no values entered (empty cells). A discussion with the client indicated that the data entry system would allow people to leave fields blank, rather than fill in a '0' or other numbers, which explained the origin of these empty cells. Given this context, an empty cell could be interpreted as implicitly indicating a '0'. One difficulty with this interpretation, however, is the inability to distinguish between a field that could in principle have had a value, and a field that was necessarily '0', because, for example, the mission in question didn't provide that type of service at all. Similarly, this interpretation would not allow for the possibility that work was carried out and simply never entered. Because of these possibilities, these empty cells were given the value 'NA' rather than '0'. This ambiguity was then taken into account in a context appropriate manner for particular data analyses.

Change to Fields: Between 2016 and 2017 a change was made to the activity categories used to collect data about case and service activities. Because the change in categories was not a straightforward splitting or combining of previous categories, the decision was made to not map old categories onto new categories, as the resulting time and case-service number entries would not be reliable. Instead, the decision was made to carry out a relatively fine grained initial analysis of the data over the 2016 – 2020 time range, and following that, where appropriate, a more coarse grained analysis of all the data, combining all activity times into a single time estimate, across the entire time range.

	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%	-	-	-	-

Figure 28: Heat map showing the percentage of days when daily log entries were made relative to the total number of calendar days.

A.2 Data Gaps

A gap in the data is defined as a date for which there is no row entry for that date in that table. for a given mission and a given month, if there are 31 days in that month, and 20 days where data has been entered into the daily log file for that mission, there are 11 data gaps in that month.

Because the current system does not enforce a data entry for every day, gaps in the data may represent one of two possibilities: a day where work was done but no log was entered, or a day where no work was done. With respect to days not entered when no work was done, these might represent days when the mission itself was open, but no cases or services were opened, or days when the mission was closed.

A summary visualization of the gaps found in the daily log data can be found in the heat map provided as Figure 28. This heat map shows the mean percentage days entered into the daily log for each mission, relative to the total possible number of days that could be entered for each mission (2016-2020 data).

A sample of the summaries of the monthly log data for each mission can be seen in Figure 29 (the file `MonthlyLogAnalysisSparkLinesPostData.xlsx` contains summary data for all missions can be found in the table). The 'Blanks' field provides information about the number of months that have no data for that mission, across the years of the dataset reviewed (2016 – 2020).

A.3 Logical Inconsistencies in the Data

Inconsistencies between Daily and Monthly Logs Relating to Total Number of Cases and Services, Daily vs Monthly: `DailyMonthlyLogComparisonPostData20212022.xlsx` contains a table comparing the summed number of case-services for daily logs and the number of case-services entered in the monthly log, for each mission.

Inconsistencies between Daily and Monthly Logs Relating to Time Worked Amount, Daily vs Monthly: Please see the file `DailyMonthlyLogComparisonPostData20212022.xlsx` for tables comparing the summed amount of time worked for daily logs (converted to hours from minutes) and the amount of time entered in the monthly log (entered in hours), for each mission.

Logical categories relating to case-service numbers and time worked are found in the file `CaseServiceTimeLogicalInconsistencesPostData20212022.xlsx`, which contains all tables comparing the number of cases and services entered with the amount of time worked for each mission, for each year, organized by category. 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 is logically inconsistent.

B Data Entry Assessment Metric

Although measures cannot be easily constructed to validate mission data, they can more readily be constructed in order to suggest when missions are, or are not, using the current data entry system properly. In particular, there are a variety of markers which indicate when a mission is using problematic data entry practices. These markers include:

- Mismatch between daily and monthly log values
- Unusual number values
- Overly consistent values
- Gaps and '0' entries

To calculate the full data entry assessment metric, missions can be rated on each individual data entry measure, and the resulting ratings weighted and summed to provide an overall data quality metric value for each mission. Unlike the previous metrics discussed in the report, the data required for this calculation is currently available and can be used to calculate this metric for each mission if desired.

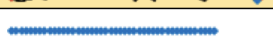
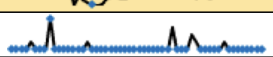
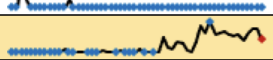
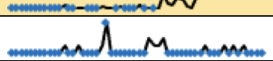
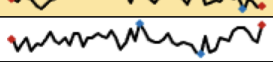
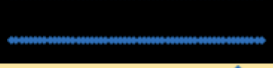
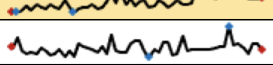
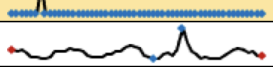
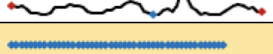
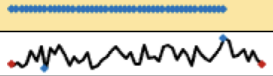
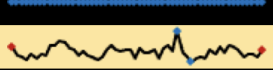
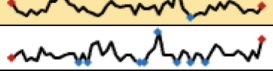
Mission	Start	Monthly Number of Cases	End	Low	High	Mean	Std Dev	Blanks	Zeros	Trend
Grand Total	19502		17265	15150	25072	19903	2612	0.0	0.0	379.2
Adriata	46		19	3	46	19	9	0.0	0.0	-1.6
Addasibaba	156		240	101	326	194	60	0.0	0.0	9.7
Adsaloma	16		11	2	76	15	15	0.0	0.0	-2.9
Afolo	3		13	0	105	9	15	0.0	0.4	-1.8
Agruma	42		50	25	91	61	16	0.0	0.0	1.2
Ahmohamabad	48		53	34	169	67	25	0.0	0.0	0.6
Airy	0		N.A.	0	0	0	0	2.2	9.8	0.0
Aitioch	56		104	34	150	73	25	0.0	0.0	4.6
Alma-Ata	0		N.A.	0	1	0	0	9.6	1.6	-0.2
Al-geris	195		189	112	544	244	91	0.0	0.0	12.4
Alia	0		N.A.	0	0	0	0	6.2	5.8	0.0
Amani	127		78	41	132	86	25	0.0	0.0	1.0
Ammertuma	0		0	0	4	0	1	0.0	10.4	0.0
Antipero	0		0	0	0	0	0	0.0	12.0	0.0
Antwoerten	0		N.A.	0	0	0	0	2.7	9.3	0.0
Appilo	0		0	0	0	0	0	0.0	12.0	0.0
Asbi	0		0	0	4	0	1	0.0	11.3	-0.1
Asneria	0		18	0	43	9	13	0.0	4.9	7.6
Atary	0		0	0	4	0	1	0.0	9.3	0.0
Atilanta	299		236	204	482	323	61	0.0	0.0	-5.0
Attila	33		50	15	53	33	9	0.0	0.0	1.1
Attor	0		0	0	0	0	0	0.0	12.0	0.0
Awez	0		0	0	0	0	0	0.0	12.0	0.0
Bakounour	2		23	1	48	18	11	0.0	0.0	6.3
Bigowan	11		9	1	33	11	6	0.0	0.0	1.2
Badigehalli	0		0	0	0	0	0	0.0	12.0	0.0
Bang Krung Thep	163		249	146	293	209	36	0.0	0.0	9.6
Bangi	0		0	0	1	0	0	0.0	11.8	0.0
Barcino	96		33	3	318	70	52	0.0	0.0	0.6
Batte	0		N.A.	0	0	0	0	1.8	10.2	0.0
Beichengshi	322		333	291	534	396	57	0.0	0.0	12.3
Beirite	606		420	321	856	553	123	0.0	0.0	18.1
Bell Town	0		0	0	0	0	0	0.0	12.0	0.0
Berograda	92		80	36	159	77	22	0.0	0.0	-0.2
Betis City	2		9	0	12	3	2	0.0	1.6	0.2
Bajo Horizonte	0		0	0	0	0	0	0.0	12.0	0.0

Figure 29: Sparklines and summary of the monthly log data for each mission (extract).