Faster Payments to Farmers: Analysis of the Direct Payments Process of EU's Agricultural Guarantee Fund

Business Process Intelligence Challenge 2018

Lalit Wangikar, Sumit Dhuwalia, Abhilasha Yadav, Bhavy Dikshit, Dikshant Yadav

Cognitio Analytics, 2 Woodland Road, Short Hills, NJ 07078, USA
{lwangikar, sdhuwalia, abyadav, bdikshit, dyadav}@cognitioanalytics.com

Abstract. The 2018 Business Process Intelligence Challenge is centered on understanding and analyzing the core process behind the direct payments program of the European Agricultural Guarantee Fund. We analyzed the data provided by process owners using a variety of process mining and analytical tools. In this report, we outline our understanding of the data and the process, present findings from the exploratory analysis of the event log data, answer specific business questions posed by the process owners, and recommend further analysis, wherever applicable. We also discuss some limitations of working with such data in absence of additional applicant specific data and a deeper understanding of the underlying business context.

Keywords: Process Mining, BPIC 2018, Event Logs, Process Discovery, Concept Drift, Predictive Modeling, Process Improvement, Operations Effectiveness

1 Introduction

Like the previous editions, the 2018 BPI Challenge provides a unique opportunity to analyze a real-world business process based on event log data, using a combination of commercial and open-source tools, to generate process related insights that can be used to drive better business outcomes.

This year, we are asked to analyze the process for handling of applications for European Union direct payments for German farmers from the European Agricultural Guarantee Fund [3]. Direct payments are a key element of the common agricultural policy (CAP) of EU that provides income support for farmers and promotes competitiveness, sustainability and environmentally-friendly farming practices. Direct payments benefit nearly 7 million farms throughout the European Union and often represent an important share of their agricultural income [1-2].

1.1 Approach and Scope

The challenge encourages participants to draw relevant process related insights that can be useful for the purposes of a real-life business improvement setting [3]. More

specifically, the process owners have formulated four business questions on the data and the participants are expected to solve one or more amongst them.

Keeping the overall objectives in mind, we analyzed the data in detail and across multiple dimensions. In doing so, we used a combination of process mining and advanced statistical techniques, leveraging commercial and open-source tools. Our overall approach has been summarized below:

- Develop thorough understanding of the provided event log data
- Understand the underlying business process, based on process discovery through process mining tools and reading documents available on EU website [2]
- Study the process maps to identify key phases of the process
- Conduct exploratory analysis to gain a deeper understanding of the process and to assess process performance
- Perform detailed analyses and develop relevant statistical models (where applicable) to answer the four business questions put forward by the process owners and suggest further analysis, as appropriate

Based on the guidelines mentioned for submissions under the professional category [3], we have attempted to answer each of the four questions posed by the process owners.

2 Data Overview

The data for BPI Challenge 2018 has been provided by the German company "data experts", located in Neubrandenburg [4]. The overall event log contains data for 43,809 direct payments applications over a period of three years from 2015 to 2017. The data consists of 2,514,266 events for these 43,809 applications. These events represent a combination of manual and automatic work steps performed, starting with receiving the application and, if all goes well, finishing with the authorization of a payment. The applications go through several process steps that establish the eligibility for direct payments and the amounts to be paid. During this process, some applications are reopened, either by the department (subprocess "Change") or due to a legal objection by the applicant (subprocess "Objection"), while some other go through inspections [3].

The process flow is organized by document types (refer Table 1), where each document type has a state (Sub Process) that allows for certain actions (Activity).

Table 1. Description of document types

Sl. no.	Document type	Sub process	Description
1	Entitlement application	Main, Objection, Change	The application document for entitlements, i.e., the right to apply for direct payments, usually created once at the beginning of a new funding period
2	Payment application	Main, Application, Objection, Change	The application document for direct payments for each year
3	Parcel document (only for 2015 cases)	Main	The document containing all parcels for which subsidies are requested
4	Department control parcels (only for 2015 and 2016 cases)	Main	The document containing the results of checks regarding the validity of parcels of a single applicant
5	Geo parcel document (replaces Parcel document since 2016 and Department control parcels since 2017)	Main, Declared, Reported	The document containing all parcels for which subsidies are requested.
6	Reference alignment	Main	The document containing the results of aligning the parcels as stated by the applicant with known reference parcels
7	Inspection	On-site, Remote	The document containing the results of on-site or remote-inspections
8	Control summary	Main	The document containing the summarized results of various checks (reference alignment, department control, inspections)

In addition to the overall event log, we were also provided individual sub logs for each of the document type. The sub logs for each document type contain records from the overall event log filtered for a given document type.

3 Process Understanding

3.1 Data Pre-Processing

The overall event log provided has well-defined columns to aid process discovery. But, to visualize the process in different ways, we processed the data further to create custom event logs tailored to specific objectives, as mentioned below:

1. **Develop a bird's eye view of the process:** Understand the process flow of the applications in terms of the most high-level dimension, i.e., document types
2. **Develop a detailed view of the process:** Leverage the combination of document type, subprocess, and activity to understand the process flow at the most granular level

Accordingly, the pre-processing done to create these custom event logs has been described below:

- **Fix potential data errors:** A total of 6 applications had timestamp originating in 2014. Considering these values to be error, we replaced the year part of the timestamp to 2015.
- **Create the most granular event description:** Each event in the log is defined as a combination of the document type, the subprocess, and the activity within the subprocess. Hence, we created a new event column that is a concatenation of the three to visualize the overall process.
- **Rename document types:** To compare process flow across years, we renamed "Parcel document" and "Department control parcels" to "Geo parcel document".

- **Create an event log at document type:** To simplify process visualization and capture just the flow among documents, we collapsed the overall event log data by combining the successive occurrences of a given document type into one process step and creating a corresponding start and end time stamp, as illustrated in Fig. 1 below. The resultant data had 623,777 records. We further use this dataset to create features such as time spent in a given document type, total waiting time between document types, etc. for predictive modeling, which we will cover in greater detail in section 6.1.

Doctype	Timestamp
Parcel document	8/3/2015 12:48
Parcel document	8/3/2015 12:49
Parcel document	8/3/2015 12:49
Control summary	8/4/2015 5:08
Department control parcels	9/15/2015 18:08
Parcel document	9/30/2015 12:16
Parcel document	9/30/2015 12:17

Doctype	Start time	End time
Parcel document	8/3/2015 12:48	8/3/2015 12:49
Control summary	8/4/2015 5:08	8/4/2015 5:08
Department control parcels	9/15/2015 18:08	9/15/2015 18:08
Parcel document	9/30/2015 12:16	9/30/2015 12:17

Fig. 1. Comparison of original and transformed sample records for application id: 6ded91cacb8fd9fc

3.2 Process Overview

Using Celonis, we discovered the overall process flow for all the applications across years, post the renaming of appropriate document types, as described in section 3.1 above.

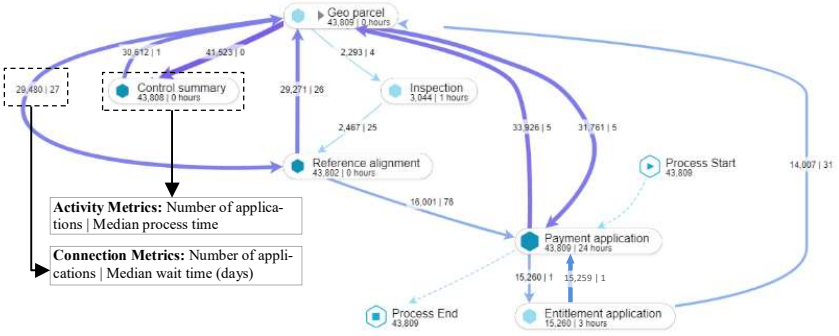


Fig. 2. Overall process flow

Given the fact that different process functions can be included within the same document type, the process flow in Fig. 2 may be misleading. For instance, the document type "Payment application" includes work flow steps - 'application for payment', 'finalization of decision', 'processing of payment', and 'reopening of an application' - because of which majority of the applications begin and end with the document type "Payment application" in Fig. 2 above. To overcome this problem, we mapped the

granular "doctype-subprocess-activity" level events into high-level work flow steps (refer Table 2).

Table 2. High-level work flow steps for each doctype-subprocess-activity

Doctype	Subprocess	Activity	Work flow step
Control summary	*	*	Update Relevant Documents
Department control parcels	*	*	Update Relevant Documents
Entitlement application	Change/Objection	*	Reopen
	Main	*	Calculation of Entitlement
Geo parcel document	Declared/Main	*	Record Parcel Details
	Reported	*	Update Relevant Documents
Inspection	*	*	Inspection
Parcel document	*	*	Record Parcel Details
Payment application	Application	abort/begin/finish payment, insert/remove document	Process payment
		mail income, mail valid	Application for payment
		begin editing, calculate, decide, finish editing, initialize, revoke decision, revoke withdrawal, withdraw	Finalize decision
	Change/Main/Objection	*	Reopen
Reference alignment	*	*	Validate Parcel Information

* - refers to all the subprocesses for a doctype and similarly all the activities for a doctype-subprocess combination

The resultant data was collapsed using the same approach described in section 3.1 and illustrated in Fig. 1. The process flow thus obtained has been shown in Fig. 3.

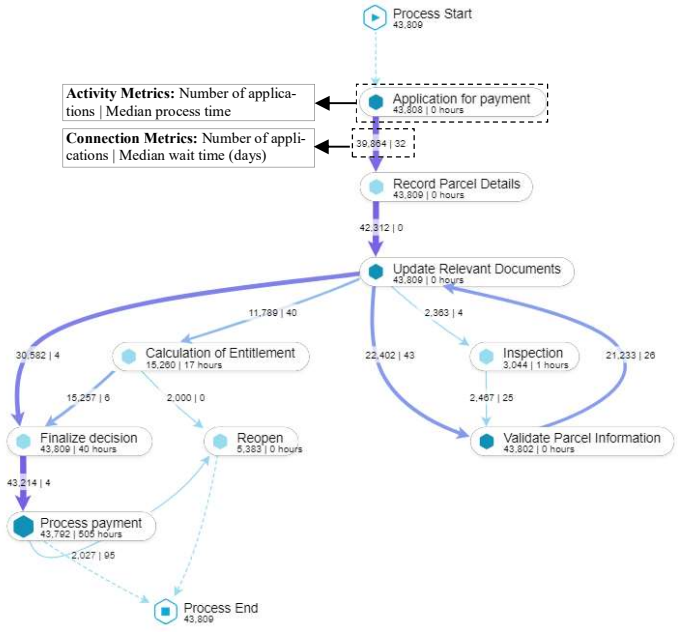


Fig. 3. Process flow with high-level work flow steps

Also, corresponding to this process, we created a generalized work flow as depicted in Fig. 4.

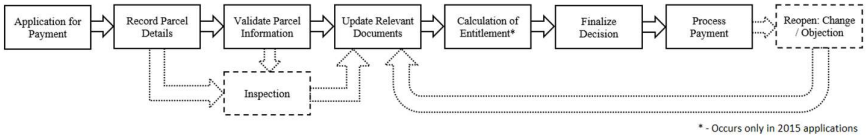


Fig. 4. Generalized work flow

For all three years, the process starts with the application for payment by an applicant. A significant proportion of the total time in an application is spent in the parcels validation process which includes recording the parcel details, validating the parcel information, and updating the result of the verification in appropriate document type (refer Fig. 5 below). During the parcel validation process, a few applications may also be selected for inspection. Compared to that in 2016 and 2017, the validation step is followed by a slightly different process in 2015. In 2015, the first year of implementation of the basic payment scheme, eligible farmers were allocated payment entitlements [1]. Therefore, in 2015, the validation step is followed by the calculation of entitlement and then by the decision finalization and the actual payment process step. Whereas, in most of the applications, the validation step is directly followed by the decision finalization and the payment process step in 2016 and 2017. Also, the processing of the payments takes around 505 hours, i.e., 21 days, as can be seen in Fig. 3 above.

Some applications may also be reopened, either by the department or due to a legal objection by the applicant, represented by "Reopen: Change / Objection" in Fig. 4 above.

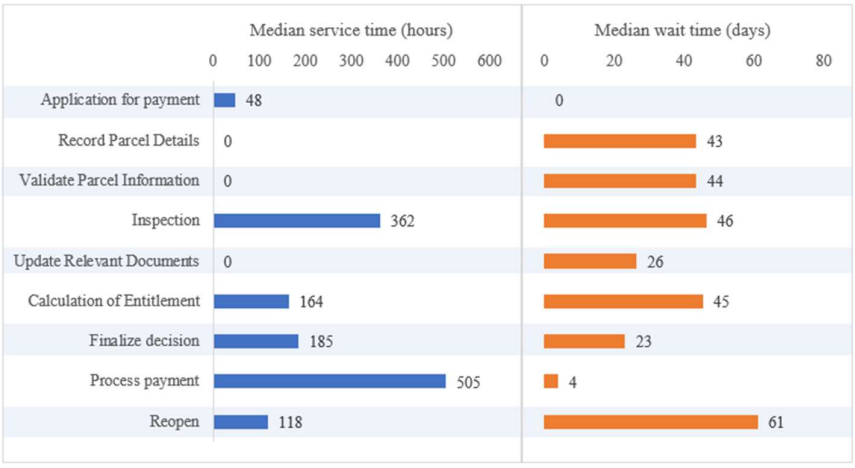


Fig. 5. Distribution of service and wait time across work flow steps

4 Exploratory Data Analysis

4.1 Distribution of Applications

Under the Basic Payment Scheme (BPS) that came into effect in 2015, farmers were allocated payment entitlements in the first year of application and the entitlements are subsequently activated each year by the farmers [5]. We studied the volume of applications received each year and found it to be relatively constant, as shown in Table 3 below.

Table 3. Volume of applications by year

Application year	# Applications
2015	14,750
2016	14,552
2017	14,507

Also, to validate our assumption that the mix of applicants across years is consistent, we calculated the graph below that shows the number of new applicants each year and the number of applicants that reapply in subsequent years. As observed in Fig. 6 below, 94% and 91% of all the applicants of 2015 reapply in 2016 and 2017 respectively and 93% of new applicants of 2016 reapply in 2017.

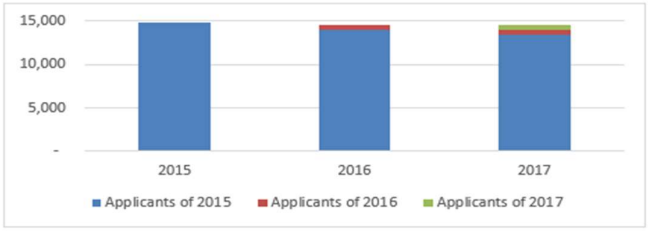


Fig. 6. Distribution of applicants by year of first application

Furthermore, we observe that most of the applications are initiated in the months of April and May each year. In 2015 and 2016, the number of applications initiated in May are more than the number of applications initiated in April, but it is not so for 2017.

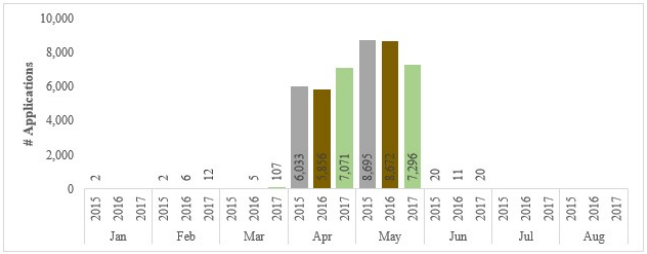


Fig. 7. Distribution of applications by application month

4.2 Distribution of Applications Across Departments

To evaluate the distribution of applications across departments, we calculated the table below that shows the proportion of applications that are processed by each department within a year.

Table 4. Distribution of applications across departments

Department	2015	2016	2017	Total
4e	4,580 (31%)	4,531 (31%)	4,528 (31%)	13,639
6b	3,793 (25%)	3,732 (25%)	3,731 (25%)	11,256
d4	1,930 (13%)	1,910 (13%)	1,904 (13%)	5,744
e7	4,447 (30%)	4,379 (30%)	4,344 (29%)	13,170
Total	14,750	14,552	14,507	43,809

As can be seen in Table 4 above, each department processes the same proportion of overall applications every year. Also, department "d4" processes far fewer applications each year compared to other departments.

4.3 Overall Duration of Applications

We next sought to determine the time it takes to complete the processing of the applications in different years. In the absence of any specific activities that define end of an application, we calculated the throughput time as the time difference between the first and the last activity of each application.

As the event data is available until 19th January 2018, we have different lengths of history available for applications filed in different years. For example, for applications that were filed in April 2017, approximately 9 months of event history is available. Whereas, for applications filed in April 2015, we have event history for over 2 years and 9 months. And some of the applications from 2015 do indeed have activities occurring as late as January 2018. Having said that, applications from 2015 seem to be most complete with only 2.8% applications having process activities during the last 3 months of the given data ("Active" applications) (refer Table 5).

Table 5. Distribution of "Active" applications

Application year	2015	2016	2017
# "Active" Applications (%)	418 (2.8%)	883 (6.1%)	14,507 (100%)

This difference in length of history of event log and lack of a clear indication that the application is permanently "closed", makes it difficult to compare the overall turnaround time across years, especially with 2017.

As illustrated in Fig. 8 below, approximately 46% and 12% of applications in 2015 have a throughput time of more than 12 months and 24 months respectively. However, only 16% of applications in 2016 have a throughput time of more than 12 months. This has resulted in the reduction of median throughput time from 448 days in 2015 to 256 days in 2016.

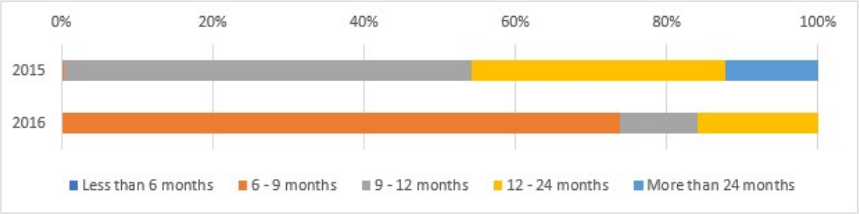


Fig. 8. Distribution of throughput time across years

4.4 Studying the Input Effort Across Years

We calculated the distribution of applications across different buckets of the number of activities for each year, given that the number of activities in an application is usually a good measurement of the effort spent on that application.

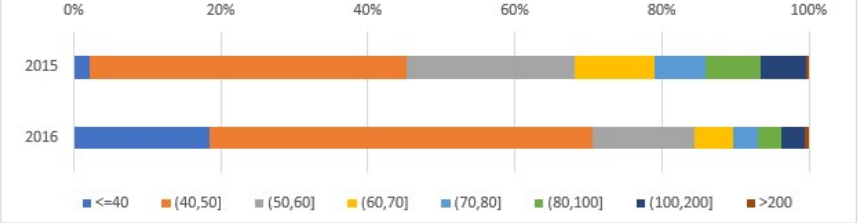


Fig. 9. Distribution of number of activities by year

As can be seen in Fig. 9 above, most of the applications see 40 to 50 activities during their processing. The applications in 2015 and 2016 take an average of 61 and 52 activities respectively. However, since all the applications of 2017 (refer Table 5) are still active, it would be premature to make any comparisons of the effort that goes into 2017 applications with those of 2015 and 2016.

4.5 Distribution of Penalties

As described on the homepage of the BPI Challenge 2018 [3], applicants may not receive the total amount that is applied for because of certain penalties. The penalties may be applied for various reasons, such as mismatch in the information provided regarding the size of the farmland, violation of cross-compliance rules, or non-compliance with the requirements of greening or young farmer clauses. To understand the nature and the

quantum of these penalties, we studied the distribution of different penalties across years.

Table 6. Distribution of different penalty types across years

Year		2015	2016	2017
Severe Penalties	B3	805	-	-
	B4	191	-	-
	B6	123	-	-
	B16	111	-	-
	JLP3	2	-	10
	C16	1	1	3
	B5	-	51	105
	B5F	-	-	9
	BGK	-	119	174
	BGKV	-	-	15
	BGP	-	-	372
	V5	-	1	-
Non-Severe Penalties	B2	7,967	2,058	1,502
	GP1	795	614	916
	CC	528	411	404
	AUVP	367	251	211
	ABP	248	245	211
	AGP	244	247	211
	C4	192	156	-
	JLP1	99	45	25
	JLP2	32	8	5
	AJLP	22	20	19
	C9	17	5	4
	AVBP	1	1	-
	AVGP	1	1	-
	AVJLP	1	2	-
	AVUVP	-	2	-
	JLP5	-	11	14
	JLP6	-	58	66
	JLP7	-	-	1
Total # Applications*		9,388 (64%)	3,368 (23%)	2,967 (20%)

* - An application may have multiple penalties, hence # applications is not a sum of the respective columns

As can be seen in Table 6 above, in total, there are 30 different penalty types, out of which 12 penalties are tagged as severe by the process owners. It is interesting to note that some of the most frequently occurring severe penalties of 2015 do not occur in 2016 and 2017, whereas the commonly occurring severe penalties of 2016 and 2017 do not occur in 2015, although the same cannot be said for non-severe penalties.

We also observe (refer Table 6 above) that 64% of all applications in 2015 had at least one penalty, whereas only 23% and 20% of applications in 2016 and 2017 respectively had so.

4.6 Summary

Overall, findings from the exploratory analysis seem to indicate that the process has gotten more effective and streamlined over the years.

5 Limitations

Understanding the direct payments process highlighted in this challenge and answering the corresponding business questions posed by the process owners requires a deep understanding of each step of the overall work flow. We have highlighted below the major challenges that we faced during our analysis:

1. **Data and business understanding limitation:** We do not have all the data about applications nor do we have a detailed understanding of the business rules and logic behind these payments. The process is also known in abstract, further details about what happens in each document type would aid in developing more complete solutions.
2. **Complete process lifecycle is available only for one year:** As discussed in section 4.3 above, only applications of 2015 seem to have completed their entire process lifecycle. For a process that is relatively new [1] and unstable, it is premature to draw concrete conclusions based on just one year of process data.

6 Responses to Questions

6.1 Q1: Undesired Outcomes

Problem Statement

The process owners have described two "undesired outcomes" that can occur in processing these applications:

- Undesired outcome 1 (UD1): the payment is late
- Undesired outcome 2 (UD2): the application needs to be reopened

The process owners would like to detect such applications as early as possible, ideally, before a decision is made for an application.

Approach

We began our analysis by first developing clear rules for identifying undesired outcome and then tagged each application that experienced one or both of these outcomes. We then quantified the magnitude of the problem and created a predictive modeling framework to detect applications that experience such outcomes, as early as possible in the process lifecycle.

Identifying Applications with Undesired Outcomes

i. Undesired Outcome 1: late payments

As per process owners, a payment can be considered timely, if there has been a "begin payment" activity by the end of the year that was not eventually followed by "abort payment". We test for this by checking if the last 'begin payment' associated with the

subprocess 'Application' occurs before the end of the year and is not eventually followed by an 'abort payment'. Applications meeting this criterion are tagged as those with timely payment and all the other applications are tagged as the ones which had a late payment (UD1). Table 7 below shows incidence of late payments by year of application.

Table 7. Number of applications with undesired outcome 1: Late payment

Application year	# Applications	# Applications with timely payment	# Applications with late payment	Event Rate (UD1) %
2015	14,750	14,654	96	0.65%
2016	14,552	14,480	72	0.49%
2017	14,507	14,501	6	0.04%

ii. Undesired Outcome 2: the application needs to be reopened

Applications can be reopened either by department (subprocess 'Change') or due to a legal objection by applicant (subprocess 'Objection'). We identified applications that needed to be reopened based on occurrence of activities under the subprocesses 'Change' or 'Objection' and tagged them as ones which had this undesired outcome (UD2). Table 8 below shows incidence rate of this outcome by application year. It also shows median time to start of reopening of the application from time of the application start.

Table 8. Number of applications with undesired outcome 2: Reopened applications

Application year	# Applications	# Applications with Change/Objection	Event Rate (UD2) %	Mean Time to UD2 (days)	Median Time to UD2 (days)
2015	14,750	3,081	21%	630	667
2016	14,552	1,816	12%	424	405
2017	14,507	81	1%	219	226

The event rate (UD2) decreases as we go from 2015 to 2017. As discussed in section 4.3, we have varying lengths of application history available for these three years. As such, for applications starting in 2016 and 2017, we may still see more applications reopen as time goes by, more so for 2017. Based on this, we decided not to use data for 2017 applications in developing a model for predicting this specific undesired outcome.

Predictive Modeling- Detecting Applications with Undesired Outcomes

Prediction Method:

We created binomial logistic models at different time intervals to detect undesired outcomes, either UD1 or UD2.

Approach:

Based on business requirements to score applications before the decision point and our observation that minimum time across all applications from process start to the decision point, i.e., 'Payment Application – Application – decide', is 180 days, we decided to use all the process related events that happen for the applications within the first 180 days of process start. Also, since the process owners want to detect such applications as early as possible, we chose to build models to make predictions at different time intervals during the process life cycle. Specifically, we built three models:

- On the day of the application start, or day 0
- 90 days from the start of the application
- 180 days from the start of the application

Each of these models uses all the process related variables available to be used at the time of prediction. Accordingly, the model for day 0 prediction uses only application characteristics known at the time of application initiation. In 'Model Training and Result'- section 6.1, we discuss how the information gain from day 0 to day 180, resulting from process related events, adds value to the model performance.

To see the impact of data from previous years' experience for the same applicant on current year predictions for the applications, we created two sets of models - one with data from previous and current year and the other without the data from the previous year. For our modeling exercise, we chose the applications which were initiated in 2016 (14,552 applications) and created 6 different models, as described in Table 9 below:

Table 9. Different types of models

Scoring Day / With previous year data	0	90	180
N	Model 1	Model 2	Model 3
Y	Model 4	Model 5	Model 6

Feature Creation:

The process owners have divided the application level attributes into "raw" and "derived" attributes. They have mentioned that the derived attributes with suffix "0" belong to the "Application" subprocess and are available at the time of the decision [6]. It's unclear to us as to when other derived attributes are recorded (and therefore available to be used for prediction) during the application lifecycle. Hence, we did not use the derived attributes from the current year that did not have a suffix "0" in developing prediction models.

In the feature description below, the features for the current year have a prefix 'CYD', while features created using data from an applicant's previous year's application have a prefix 'PYDI'. The data from the previous year's application has been filtered to

include only the data available at the time of prediction. For instance, for the 90 day model, the data from the previous year has been filtered till the 90th day from the date of initiation of the current year application.

Different features created using information available from current year and previous year's application of an applicant, can be categorized as:

1. **Application related features:** Raw and transformed features created using 'raw' and 'derived' attributes of an application. We describe a few of these features in Table 10 below:

Table 10. Some application related features

	Variable name	Description
Raw	CYD_department	Department of an application
	PYDI_department	
	CYD_number of parcels	Number of parcels in an application
Transformed	CYD_area per parcel	Area per parcel in an application
	CYD_log amount applied 0	Log normalized amount applied 0 of an application
	PYDI_number of penalties	Total number of penalties applied in an application

2. **Process features:** Features capturing process information in the application lifecycle. Some of these features have been shown in Table 11 below:

Table 11. Some process related features

	Variable name	Description
Transformed	CYD_Payment Application count	Count of doctype 'Payment Application' in an application
	CYD_Payment Application flag	Flag if doctype 'Payment Application' has occurred in an application
	CYD_time since last activity	Number of days passed till date, since the last activity occurred in an application
	CYD_intime in Payment Application	Total time taken in 'Payment Application' calculated using event log at doctype level as mentioned in 'Preprocessing-4' - section 3.1

Model Training and Result:

Following are the basic steps that we performed while building each of the six models:

- We created datasets containing features till the appropriate day, i.e., 0, 90, or 180 days.
- The data was split into train and test in the ratio of 70:30, keeping consistent, the application ids in train and test for each dataset.
- We performed the necessary data processing steps such as outlier treatment, missing value imputation, etc.

- We also performed various iterations for variable reduction using correlation and VIF [10], which checks for multicollinearity among the predictors. This was followed by stepwise regression for selecting significant variables.

We compared the different models using AUC (area under the ROC curve [11]) and decile-wise precision and recall. As can be seen in Fig. 10 below, the model performance is consistent across the train and the test datasets for each of the different models.

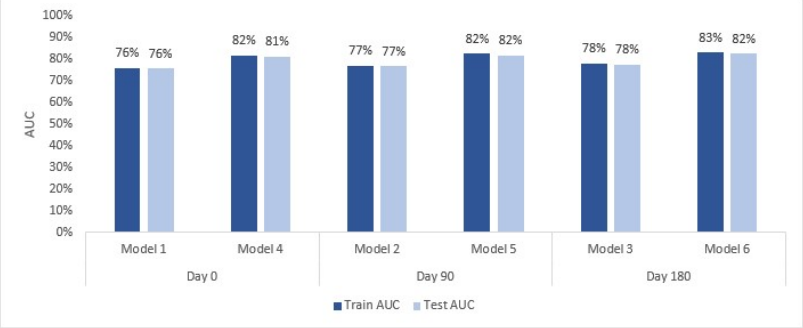
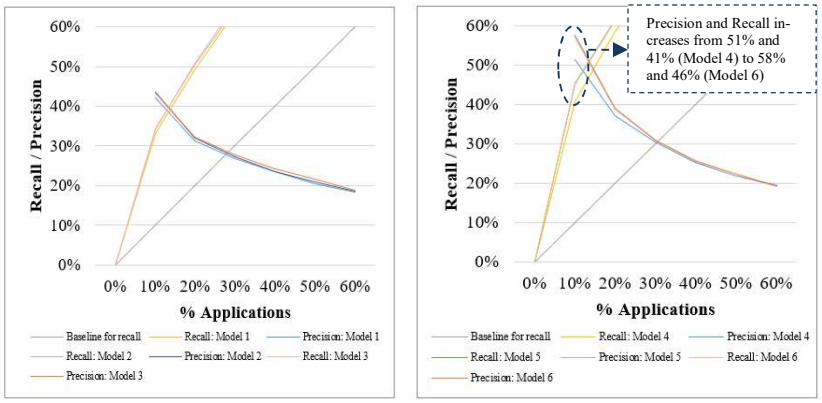


Fig. 10. AUC comparison for all models



(a) Models with no data from prior year (b) Models with data from prior year

Fig. 11. Decile-wise precision and recall of different models[1]

Comparison of Models Across Scoring Days:

We observe that both the types of models, i.e., with and without previous year data included see an improvement in model performance (refer Fig. 10, Fig. 11 (a) and Fig. 11 (b)) as we go from day 0 to day 180. While the improvement in AUC is minimal, the prediction performance, as shown by precision and recall across deciles, show a

[1] Precision and recall is shown only for test data as the performance on the train data is similar

marked improvement as we predict later in the process lifecycle. For example, Model 3 - day 180 model without prior year's data - performs better than Model 1 - day 0 model with no data from prior year.

Model 1 is basically characterized by four variables (refer Table 12):

1. Number of parcels: The more the number of parcels, higher the chances of the application meeting an undesired outcome.
2. Amount applied 0: Higher the Amount applied 0 - the initial amount applied for in an application - and Amount applied 0 per unit area, lower the chances of the application meeting an undesired outcome.
3. Case department: The case department being 'd4' or '4e' adds to the probability of an application meeting an undesired outcome.
4. Application initiation month: The applications initiated in the month of May are more prone to having an undesired outcome than the others.

Table 12. Model 1 – Final model features

Sl. no.	Variable name	Estimate	z value	p value
1	CYD_number of parcels	0.0426	25.01	0
2	CYD_avg amount applied0 per parcel	0.0004	10.3	0
3	CYD_avg amount applied0 per unit area	-0.0088	-9.44	0
4	CYD_department -d4	0.7083	7.97	0
5	CYD_log amount applied 0	-0.0946	-6.7	0
6	CYD_department -4e	0.3427	4.73	0
7	CYD_application initiation month -May	0.2502	3.71	0
8	Intercept	-0.0566	-0.2	0.84

As we predict later in the life-cycle, process related variables also appear as important predictors. For example, in the 180 day model with no previous year data (Model 3), process related features, such as count and flag of activities at various levels, and percentage of successful activities in an application, appear significant (refer Table 13). The percentage of successful activities (Feature #7, Table 13) having a negative estimate indicates that applications with positive movement forward in the process are less likely to see occurrence of undesired outcomes.

Table 13. Model 3 – Process related features

Sl. no.	Variable name	Estimate	z value	p value
1	CYD_'Entitlement application - Main - mail valid' count	0.7123	8.4	0
2	CYD_'Control summary - Main - save' flag	0.4922	4.26	0
3	CYD_'Geo parcel document - Main - save' flag	0.2818	3.27	0
4	CYD_'Inspection - Remote - finish editing' count	0.4615	3.16	0
5	CYD_'Inspection - On-Site - prepare offline' count	0.4406	3.11	0
6	CYD_'Geo parcel document - Declared - begin editing' count	0.1163	2.88	0
7	CYD_percentage of successful activties	-4.8888	-2.39	0.02
8	CYD_'calculate' flag	-0.3293	-2.22	0.03

Comparison of Models With and Without Previous Year Data:

The addition of previous years' data provides better predictive capability to our models, as we see a 6 percentage points increase (refer Fig. 10) in model AUC from 76% in Model 1 (day 0 model without previous year information) to 82% in Model 4 (day 0

model with previous year information). We see that, revoking decision in previous year application for an applicant (Feature #9, Table 14) results in higher chances for the current year application to meet an undesired outcome. In addition, the derived attributes from previous year application namely, penalty amount 0, penalties AUVP and B2 (Features #3, #6 and #7, Table 14) are significant predictors in Model 4. However, interpreting the importance of each significant variable in the model requires a thorough business understanding. For example, understanding the difference between B2 penalty and other 'B' group penalties would help us explain the importance of B2 penalty in predicting undesired outcomes.

Table 14. Model 4 – Final model features

Sl. no.	Variable name	Estimate	z value	p value
1	CYD_avg amount applied0 per unit area	-0.0118	-11.34	0
2	PYDI_'Payment application - Application - begin editing' count	-1.1297	-6.03	0
3	PYDI_penalty amount 0	0.0002	5.35	0
4	CYD_department -d4	0.397	3.49	0
5	PYDI_intime in Payment Application	-0.0049	-3.41	0
6	PYDI_penalty -AUVP	-0.8843	-2.76	0.01
7	PYDI_penalty -B2	0.2272	2.75	0.01
8	PYDI_'remove document' flag	-0.6923	-2.67	0.01
9	PYDI_'revoke decision' flag	0.2492	2.59	0.01
10	CYD_department -4e	0.2621	2.57	0.01
11	PYDI_time since last activity	-0.0038	-2.24	0.03

Conclusion

We see that undesired outcomes can be predicted with good certainty as early as when an application is submitted. The quality of prediction improves substantially when we include data from prior year and as we go forward in the process lifecycle (i.e. Day 180 model performs better than Day 90 and so on). Such prediction can be used to develop targeted interventions early in the application lifecycle to reduce effort, and to improve timely performance. One can also imagine such changes creating a better customer experience for the farmers for whom these payments are an important source of livelihood.

Challenges and Future Scope of Work

Apart from the limitations highlighted in section 5, we also faced challenges while creating process related features for predictive modeling. Incomplete data understanding, for example - not knowing when the derived variables are recorded in an application lifecycle, prevented us from creating additional features that would have otherwise improved model performance. For instance, process feature like 'Number of penalties' for an application till the day the model is scored would provide important additional information about the application, which cannot be computed currently.

Although the current solution scores applications only at three time intervals, as a next step, one can build models at shorter time intervals, with inputs from business, and optimize for the relevant parameter. One can also use more advanced machine learning algorithms such as Recurrent Neural Network that can use the sequence of activities to

predict an outcome or to predict the most likely next work-step for an application. Such predictions can help plan work better and deliver a better customer experience.

6.2 Q2: Prediction of Penalties

Problem Statement

Every year a certain number of applications are selected for inspection - either randomly, manually or based on a risk assessment. As a result, applications might receive one or more penalties, some of which are considered severe. The process owners want the participants to draw a sample of size ~5% with a better recall in selecting applications with severe penalties.

They also want us to assess, statistically, whether the current year's risk assessment of an application is independent of the applicant's previous year application.

Approach

To answer this question, we performed the following steps:

- Quantify the incidence of severe penalties: We identified all the applications with severe penalties across years
- Understand the current inspection process and outcomes
- Decide the methodology to be used for predictive modeling: Identify the statistical techniques to be used, define the train and test populations, establish the time to make the prediction, and identify the predictors to be used in the model
- Build and compare the performance of different predictive models
- Understand relationship of current year's risk assessment with previous year information
- Suggest a way of selecting 5% applications for inspection

Incidence of Severe Penalties

An application is considered to have a severe penalty if it has one or more of the penalties mentioned in Table 6. Table 15 below shows the incidence of severe penalties by year of application.

Table 15. Year-wise distribution of applications with severe penalties

Application year	# Applications	# Severe penalty applications	Event rate (%)
2015	14,750	1,189	8.06%
2016	14,552	171	1.18%
2017	14,507	643	4.43%

Current Inspection Process and Outcomes

Not all the applications with severe penalties have gone through an inspection. We do not completely understand the process behind identifying such applications, so we decided not to include non-inspected applications in our model development process.

Table 16. Year-wise distribution of inspected applications with severe penalties

Application year	# Inspected applications	# Severe penalty applications	Event Rate (%)
2015	1,047	188	17.96%
2016	986	78	7.91%
2017	1,011	240	23.74%

Predictive Modeling- Detecting Applications with Severe Penalties

Population Selection:

- We used the applications of 2016 to train our models, primarily so that we can establish the impact of previous year's information on current year's risk assessment and so that we can perform an out-of-time validation on 2017 applications.
- We only used applications that were inspected (based on any of the selection criteria: random, manual or risk assessment) in 2016 to build our models because we consistently observe a difference in the event rate for applications that were inspected versus those that were not, across the years (refer Table 17), which essentially means that we do not capture all the applications, from the non-inspected pool, that would have gotten a severe penalty had they been inspected.

Table 17. Distribution of applications with severe penalties by year and inspection criteria

Application year	Inspection criteria	# Applications	# Severe penalty applications	Event rate (%)
2015	Not inspected	13,703	1,001	7.3%
	Randomly selected only	273	49	17.9%
	Other selection criteria	774	139	18.0%
2016	Not inspected	13,566	93	0.7%
	Randomly selected only	343	20	5.8%
	Other selection criteria	643	58	9.0%
2017	Not inspected	13,496	403	3.0%
	Randomly selected only	320	55	17.2%
	Other selection criteria	691	185	26.8%

Model Selection:

- To establish the impact of previous year information on current year's risk assessment, we built two binomial logistic models – one that includes information from previous year and the other without information from the previous year. The impact, in general, can be quantified in two ways, i.e., if:
    1. Features created using previous year information comes out to be significant in the model
    2. Performance of the model that includes information from previous year is better than that of the model without previous year information
- Also, inspection of applications begins after a certain date every year. Hence, we used all the process related information available till this date to train and score the applications.

Feature Creation:

We used similar methodology to create application level features as mentioned in "Feature Creation" of Section 6.1. "CYD" and "PYDI" prefixes are used for current and previous year's features respectively.

Further, process features were created only till one day prior to the day on which inspections start each year [3]. For example, to calculate process features such as "CYD_Payment Application count" on 2016 data, activities corresponding to "Payment Application" were counted only until 28th June 2016, i.e., one day prior to 29th, June 2016, the date on which inspections began in 2016.

Data Preparation and Model Building

We followed the same methodology as mentioned in "Model Training and Results" of section 6.1 for data preparation step of both the models, i.e., models with and without information from the previous year.

Next, we split the data randomly into train and test in the ratio 80:20 and performed several iterations, including VIF [10] and stepwise reduction, on the train data to get to the most significant variables of each model. Given below is the list of statistically significant variables for the train data of model that includes information from the previous year.

Table 18. Significant variables in the model with previous year information

Sl. no.	Variable name	Estimate	p value	z value
1	Intercept	-1.5858	0.0000	-5.6525
2	PYDI_Payment Actual1	-0.0025	0.0000	-5.3014
3	CYD_Avg area per parcel	-0.5412	0.0000	-4.8546
4	PYDI_Diff b/w amount applied0 payment actual0	0.0004	0.0000	4.5212
5	CYD_IndicatorGeo Parcel Doctype-Not Occurred	1.1390	0.0078	2.6615
6	PYDI_Penalty Amount0 Bin (198-270)	1.3461	0.0084	2.6341

We can see from the table above that out of the five most significant variables, three (Feature #2, #4 and #6, Table 18) are from the previous year information of the applicant. Hence, it seems that application characteristics of previous year do impact current year's risk assessment. Apart from previous year information, current year's process features also seem to impact an application's risk assessment. For example, if no activities related to Geo Parcel doctype occur until the scoring day (Feature #5, Table 18), the probability of an application to have a severe penalty increases.

Further, in the model without previous year variables, a mix of application (Feature #1, #2, Table 19) and process related (Feature #3, #4, Table 19) features came out as significant predictors.

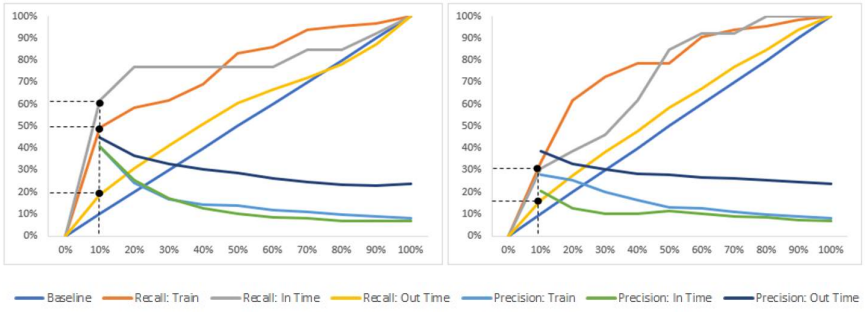
Table 19. Significant variables in the model without previous year information

Sl. no.	Variable name	Estimate	p value	z value
1	CYD_Avg area per parcel	-0.6635	0.0000	-4.6277
2	CYD_Amount Applied0	0.0000	0.0016	3.1638
3	CYD_Indicator time since last activity Bin (5.29-5.85)	1.3021	0.0054	2.7836
4	CYD_Geo parcel document - Declared - begin editing count	-0.9310	0.0090	-2.6138
5	Intercept	-0.7624	0.1038	-1.6268

Validation:

Model with information from previous year: We validated our model output on the test data, i.e., 20% data of 2016 (in-time) as well as on the inspected applications of 2017 (out-time). We calculated the decile-wise recall and precision of the model to quantify its performance. Given below (refer Fig. 12 (a)) is the cumulative gains chart for train, in-time and out-time validation.

We can observe that more than 50% and 60% events are captured in the first decile of the train and in-time data respectively. However, we observe a drop in performance on the out-time validation with only 20% events being captured in the first decile. Also, the AUC for train and in-time data is consistent at 77% while it drops to 58% for out-time data. The fact that the model performs great on in-time validation but fails to show a similar performance on out-time validation means that there must have been important process related changes that are influencing the outcomes of 2017 applications. Further inputs from the process owners would be required to refine the model and incorporate relevant process related features.



(a) Model with previous year information (b) Model without previous year information

Fig 12. Cumulative gains chart

Model without information from previous year: We also validated the performance of the model without previous year information on the in-time as well as on the out-time data. The cumulative gains chart corresponding to the train, in-time and the out-time data for the same is provided in Fig. 12 (b).

We can see that only ~31% events are captured in the first decile of the in-time data compared to 62% for the model with previous year information. Also, the AUC for this model on in-time data is 71% compared to 77% for the other model with previous year information. Hence, we can conclude that information from previous year seems to impact the risk assessment of current year's application.

Conclusion

- Our model built on inspected applications and using prior year information captures more than 50% of the events using only 10% of the applications. This allows us to recommend using this model to select 5% applications for inspection in the upcoming year. However, we cannot conclusively say that we would achieve a similar performance because of differences in the population on which the model is trained (inspected applications only) and the population which would be scored (all applications).
- To overcome this shortcoming, we recommend that process owners design a controlled experiment selecting applications for inspection using the proposed model as well as the current risk based method. They should also continue to select applications for inspection randomly.
- Insights from the above experiment would allow us to assess the performance of the proposed model on the overall population, and to further refine the model, making it more robust.

Challenges and Future Scope of Work

In addition to the data and process related challenges mentioned in "Challenges and Future Scope of Work" of section 6.1, there was an ambiguity in the selection method of an application for inspection, which made it difficult to select relevant population for modeling. For example, an application could have been selected randomly as well as on the basis of risk assessment. Also, an applicant may be selected for inspection for another type of application (other than direct payments) and then a joint inspection may be done for the different applications of that applicant [9], but we had no information in the data to separate the inspections carried out only for direct payments.
Also, additional information about the meaning of different penalties and how they work would help us to further refine our predictors, thereby improving model's performance.

6.3 Q3: Difference in Process Across Departments

Problem Statement

It might be so that the departments have implemented their processes differently. The business would like to characterize these differences across departments, if any, and would also be interested to see whether there is a relation between the way a department executes its process and the problems described in questions 1 (undesired outcomes, refer section 6.1) and 2 (severe penalties, refer section 6.2).

Introduction

In the dataset provided, there are a total of four different departments. In section 4.2, we have already shown the distribution of applications across departments and years. It is observed that every department processes the same proportion of applications every year. But within a given year, the number of applications that are processed by different departments vary. In particular, department 'd4' processes substantially fewer applications than the other departments.

Approach

For this question, we analyzed the data across departments, but separately for each year. We did so because some of the document types have changed / been replaced across years, and as we will see in section 6.4, the process has changed (drifted) across years as well.

We used a three-step framework for our analysis, as described below:

a) **Characterize process and establish differences**: We did some baseline analysis to study the process execution by different departments. We also defined and studied some process characterization metrics to quantify process differences.

b) **Analyze application characteristics**: We sought to analyze whether the differences identified in the baseline analysis were a result of the differences in the nature of applications processed by different departments or not.

c) **Outcome characterization**: We further identified the correlation between process outcomes and the differences in process characterization metrics using some important KPIs. We also sought to answer if there is indeed a relation between the different processes and the problems described in the first two questions.

Characterize Process and Establish Differences

To compare processes across different departments, we defined and analyzed some important process execution related metrics, as described below:

Number of Events per Application:

As mentioned in section 4.3, the applications for 2017 are not yet complete. Hence, we refrained ourselves from including the 2017 applications in this analysis. For 2015 and 2016, we observe that department 'd4' consistently differs in the frequency of events, and the mix of events across document types, among all the departments (refer Fig. 13).

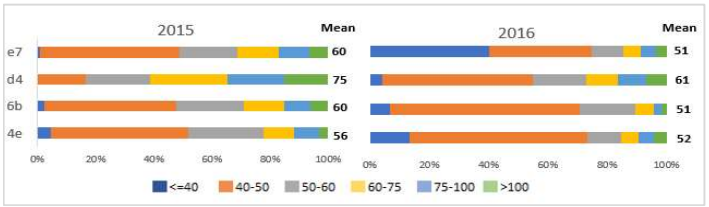


Fig. 13. Distribution of events per application across departments year-by-year

Also, ‘d4’ applications have more ‘Payment application’ related events in both the years. In 2016, we see ‘d4’ taking the most number of events for ‘Inspection’ as well (refer Fig. 14).

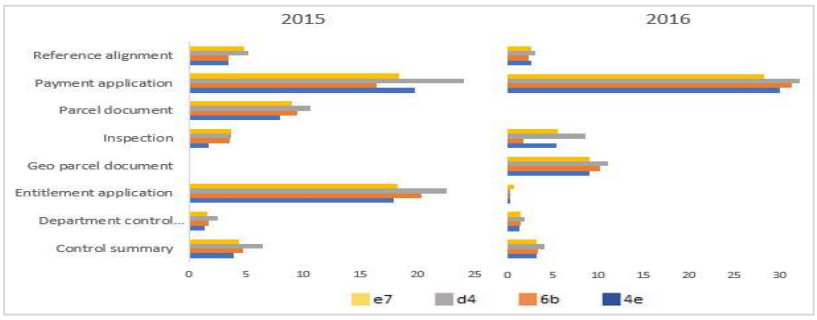


Fig. 14. Mix of events per application across document types and departments year-by-year

Distribution of Applications by Resources Across Department and Year:

We again found differences in process execution across departments in terms of resource utilization. Department ‘4e’ is found to use fewer resources than the others (refer Fig. 15).

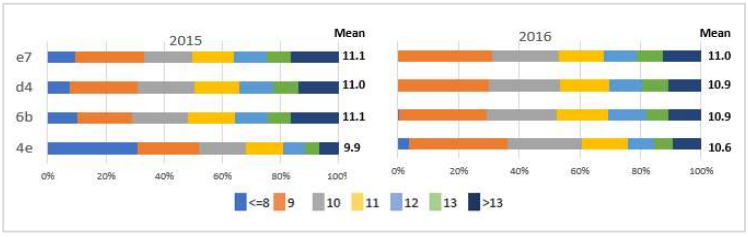


Fig. 15. Distribution of applications by resources across departments year-by-year

Analyze Application Characteristics

The metrics used to assess the characteristics of applications across departments are as described below:

Area and Parcel Information per Application:

Departments ‘d4’ and ‘6b’ have higher average area per application and area per parcel. Also, ‘4e’ has the lowest values for these metrics (refer Table 20). This correlates with ‘d4’ having highest number of events and ‘4e’ utilizing least resources, as shown in the previous section.

Table 20. Distribution of average number of parcels / area per application / area per parcel across departments and years

	Avg no of parcels				Avg area per application				Avg area per parcel			
Department	4e	6b	d4	e7	4e	6b	d4	e7	4e	6b	d4	e7
2015	17.5	18.0	17.1	19.3	58.9	75.7	75.3	65.3	3.4	4.2	4.4	3.4
2016	17.8	18.5	17.3	19.8	59.2	77.0	75.5	66.0	3.3	4.2	4.4	3.3
2017	18.0	18.7	17.6	19.8	59.1	77.0	75.6	66.3	3.3	4.1	4.3	3.4

Percentage of Applications with Small/Young Farmer:

We observe that there is no significant difference in the percentage of such applications across departments and years (refer Table 21).

Table 21. Distribution of percentage of small / young farmer applications across departments and years

	Small farmer				Young farmer			
Department	4e	6b	d4	e7	4e	6b	d4	e7
2015	5%	6%	6%	5%	8%	7%	6%	8%
2016	4%	5%	5%	4%	9%	7%	8%	9%
2017	4%	4%	5%	3%	9%	8%	9%	10%

From this analysis, we conclude that only some of the application metrics such as average area per application and average area per parcel seem to be correlated with the differences in the process characterization metrics such as number of events and resource utilized, while others do not. Hence, definitive conclusions on root causes for these differences can only be drawn based on additional data about applications and about the business rules that drive the process.

Outcome Characterization

To study how process outcome is correlated to the differences in process characterization metrics, we analyzed the following:

Throughput Time:

Department 'd4' in general takes more time than others to process the applications (refer Fig. 16). This is most evident for 2015, where it takes 565 days on average and is also somewhat observable for 2016. This seems correlated to the analysis done previously where 'd4' was observed to have more events than the other departments.

Fig. 16. Distribution of applications by throughput time (in days) across departments by year

Time to Decision:

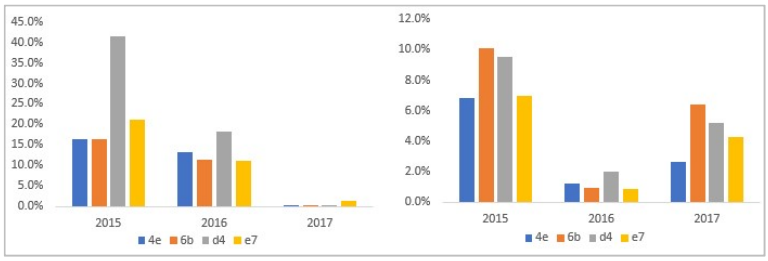
Interestingly, ‘d4’ was the fastest in taking decisions regarding its applications in 2015 (refer Table 22), but it took the most amount of time on average to completely process them.

Table 22. Distribution of the average time (in days) to finalize decision across departments and years

Department	2015	2016	2017
4e	218	223	229
6b	217	221	226
d4	216	222	227
e7	219	224	228

Relation with Undesired Outcomes and Severe Penalties:

We found that not only process outcomes like throughput time, but also undesired outcomes (Q1, section 6.1) and severe penalties (Q2, section 6.2) are correlated to the differences in process execution across departments. Department ‘d4’ is observed to have substantially high percentage of applications (40%) with undesired outcomes, as shown in Fig.17 (a) (correlated with high number of events). Department ‘4e’ is consistently seen to have fewer applications with severe penalties as compared to other departments (Fig.17 (b)). This seems to be correlated with our analysis demonstrating fewer resources being utilized by ‘4e’. Additionally, departments ‘d4’ and ‘6b’ have higher severe penalties, which correlates to these departments having higher average area per application and area per parcel.



(a) (b)

Fig. 17. (a) Percentage of applications that see undesired outcomes by departments across years as per Q1 (b) Percentage of applications that see severe penalties by departments across years as per Q2

Process Flow Changes:

It is now established that the process of ‘d4’, for instance, differs compared to other departments at least in terms of number of events, throughput time and undesired outcomes. We analyzed these differences between the process execution of ‘d4’ and that of other departments using the conformance checking feature available in Myinvenio

[7]. We performed this exercise for 2015 because the difference in 2015 is most prominent. A process model made using all the 1,920 applications for 'd4' in 2015 was compared against the reference model having 12,820 applications for all the other departments in that year.

Table 23. Process flow differences in 'd4' v/s other departments for 2015

	d4		4e, 6b and e7	
Relations	#Occurences	Occurrence per case	#Occurences	Occurrence per case
Refrence alignment to Control Summary	1,936	1.00	4,219	0.33
Control Summary to Refrence alignment	1,490	0.77	4,781	0.37
Control Summary to Department Control Parcels	913	0.47	2,473	0.19
Department Control Parcels to Control Summary	1,200	0.62	2,673	0.21

From the observations in Table 23, we infer that 'd4' is probably conducting its checks more diligently than the other departments. Higher frequency per application for all these relations means more rigorous checking being done regarding the validity of an applicant's parcels information (Reference alignment) and thus a need to summarize and record these more frequently (Control summary and Department control parcels). We also observe that in 2015, 'd4' inspected 9.14% of its applications whereas the overall inspection rate in 2015 was 7.1%.

Conclusion

We established differences in process execution characteristics across departments. Furthermore, these differences seem to be correlated to some of the application characteristics such as average area per application and area per parcel. Also, it is observed that these differences are correlated with process outcomes (throughput time), undesired outcomes (Q1) and severe penalties (Q2).

Future Scope of Work

For future, we recommend doing further analyses to try and pin-point the exact changes in the process flow, with inputs from the process owners, across departments which may be leading to a high event rate for undesired outcomes and severe penalties. This analysis can also be performed at the most granular activity level, i.e., the concatenation of three fields (Document type, Sub process, and Activity), as mentioned in section 3.1.

6.4 Q4: Differences in Process Across Years (Concept Drift)

Problem Statement

The process should be similar each but there might be some differences in process implementation across years. The business would like to characterize these differences across years, if any, in terms of 'Concept Drift'.

Introduction to Concept Drift

Any real-life process undergoes some changes over a period of time. Sometimes the changes are mandatory, like regulation or policy changes, and sometimes they can simply be a seasonal effect or a result of different people handling the same process. Due to this, a process at the beginning of the recorded period may not be the same as the process at the end. In process mining, this change during the lifecycle of a process is referred to as 'Concept Drift'.

Approach

For the scope of the analysis, we have limited ourselves to identifying instances of 'Sudden Drift' [8] in terms of 'Control Flow' changes in the data.

We have followed a three-step approach to answer the question, as explained below:

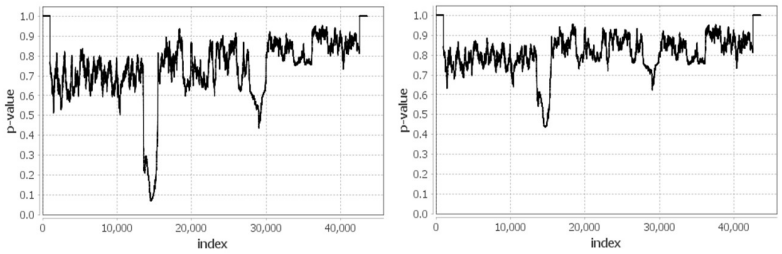
a) **Identify points of change**: We identified if there are changes in the process execution across years. We further determined the exact change points.
b) **Identify the changes**: We discovered the nature of the changes and the changes themselves.
c) **Quantify the impact on process**: We further studied the impact of the aforementioned changes on the process by defining some important KPIs.

Identification of the Points of Change

For the purpose of identifying the control flow changes in the process across years, we selected a feature, J-measure. For detailed mathematical description of the feature and the methodology used, one can refer to the research paper provided in the reference section [8].

Results

To implement our methodology using J-measure, we employed the concept drift plugin in ProM. The graphs of the p-values for the univariate KS test, used to check whether two univariate samples belong to the same distribution, were obtained as a result. We implemented this methodology on the dataset that is mentioned in section 3.1 which is at document type level only. We further filtered this dataset to have 14,500 applications from each year in order to homogenize the data quantum across years. The population size was taken to be 1,000 for comparison. Fig. 18(a) clearly shows a dip in the p-value when the applications transition from one year to another (at indices 14,500 and 29,000).



(a) (b)

Fig. 18. (a) Average significance probability (over all activity pairs) of KS-test on J-measure with original nomenclature (b) Average significance probability (over all activity pairs) of KS-test on J-measure with changed nomenclature

To understand whether this observed dip in p-value is due to a mere nomenclature change of the document types across years or is there an actual control flow change in the execution of the process as well, we homogenized the document nomenclature according to the nomenclature of 2017, as mentioned in section 3.1, and repeated the same process with the same parameters. From Fig. 18(b), it is evident that there is still a dip in p-value at indices 14,500 and 29,000, showing it is more than just the nomenclature of the documents that has changed over the years. Though, the intensity of the dip has decreased, as expected, since there is no more a difference in the names of the documents, but just the way they are executed.

Identification of the Changes

Next, we identified these changes causing the drift using the conformance checking feature present in the process mining tool Myinvenio. We used the same data as used in the section above to identify the process differences across years.

2016 vs 2015

- Since the frequency of activities corresponding to 'Entitlement application' is significantly high in 2015 [1], 'Department control parcel' which was followed by 'Entitlement application' and then by 'Payment application' in 2015 is now directly followed by 'Payment application' in 2016 (refer Fig. 19).
- In 2015, applications go from 'Parcel document' to 'Control summary' and then come back to 'Parcel document', and then finally go to 'Reference alignment'. In 2016 therefore, it is expected to see a 'Geo parcel' to 'Control summary' relation, followed again by 'Geo parcel'. This however, is not the scenario leading to control flow changes where 'Control summary' is followed directly by 'Reference alignment' in most of the applications and not by 'Geo-parcel' document. We think that this difference can be attributed to the way the two documents, 'Parcel document' and 'Geo-parcel' are executed.

- ‘Department control parcels’ to ‘Reference alignment’ is an instance of backward flow. We say this because ‘Department control parcel’ contains the results of all the validity checks of parcels, and hence must be executed only after the checks are done (‘Reference alignment’). These backward flows may lead to more rework in the applications.

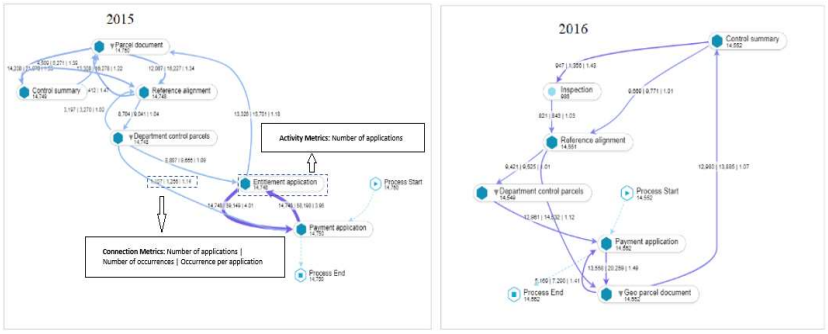


Fig. 19. Comparison of processes for 2015 and 2016

2017 vs 2016

- In 2016, reference checks were done (‘Reference alignment’) and were then recorded (‘Department Control parcels’). In 2017 however, after reference checks, the ‘Payment application’ is initialized, and then the checks are recorded (now in ‘Geo-parcel’ document). This creates a process flow ‘Reference alignment’ followed by ‘Payment application’ followed by ‘Geo parcel’ document (refer Fig. 20).
- Since ‘Department control parcels’ has been merged with ‘Geo parcel’ document, the applications that go from ‘Department control parcels’ to ‘Payment application’ in 2016, now go from ‘Geo parcel’ document to ‘Payment application’ in 2017.
- ‘Payment application’ should ideally come after the checks regarding the parcel validity (‘Reference alignment’). But we see that for 2016, around 3,000 applications go back from ‘Payment application’ to ‘Reference alignment’. We think that these connections depict backward flow in the process and increase rework.
- Like 2015, in 2017 as well, we see a ‘Geo parcel’ – ‘Control summary’ – ‘Geo parcel’ loop followed by ‘Reference alignment’. This was not present in 2016 due to which applications directly flow from ‘Geo parcel’ to ‘Control summary’ to ‘Reference alignment’.

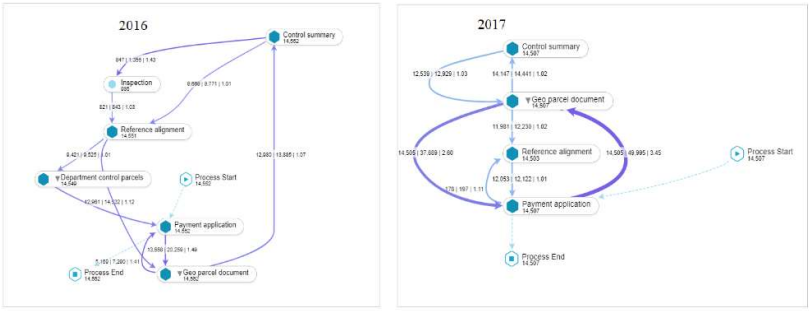


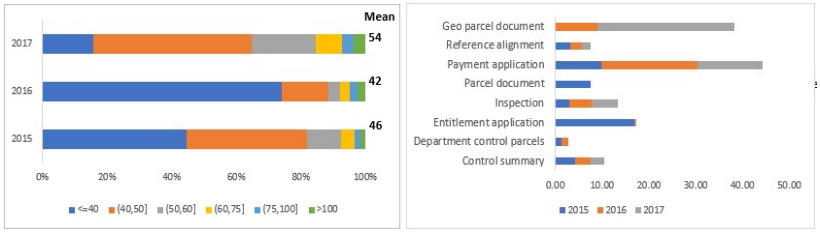
Fig. 20. Comparison of processes for 2016 and 2017

Quantification of the Impact on the Process

To assess the impact of these differences, we defined some process characterization metrics which are discussed below in detail. For this analysis, we trimmed the applications from Process Start to the first occurrence of the event 'Payment application-Application-decide', i.e., the time taken from application start to the first decision point. We did so to ensure a like to like comparison since, as mentioned earlier, the applications of 2015 have evolved for a longer time compared to the applications of 2016 and 2017 (refer section 4.3).

Distribution of Applications by Number of Events per Application from Start to First Decide Across Years

It can be seen that the number of events increases as we move ahead in the timeline (Fig.21 (a)). We also observe that the main contributors to this increase of events are the documents 'Geo parcel' document and 'Inspection' (Fig.21 (b)).



(a) (b)

Fig. 21. (a) Distribution of applications by events per application from start to decide across years (b) Distribution of events per application across document types from start to decide across years

Distribution of Applications by Resources Across Years from Start to Decide

Even though applications take more events to reach decision in 2017, they utilize fewer resources in decision making (refer Fig.22).

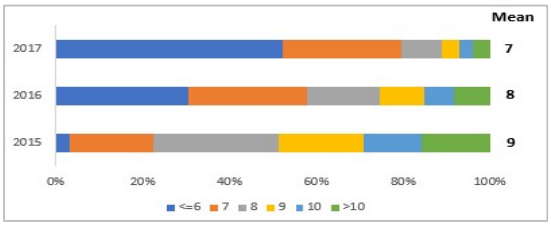
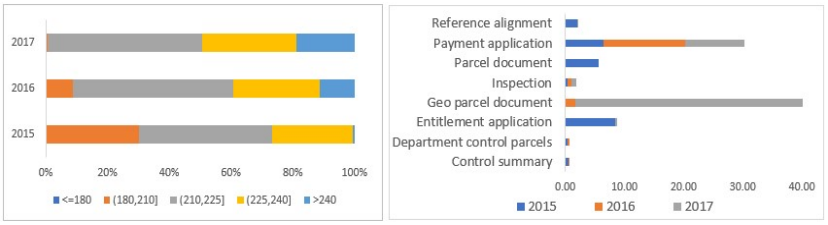


Fig. 22. Distribution of applications by resources from start to decide across years

Distribution of Throughput Time from Start to First Decide Across Years

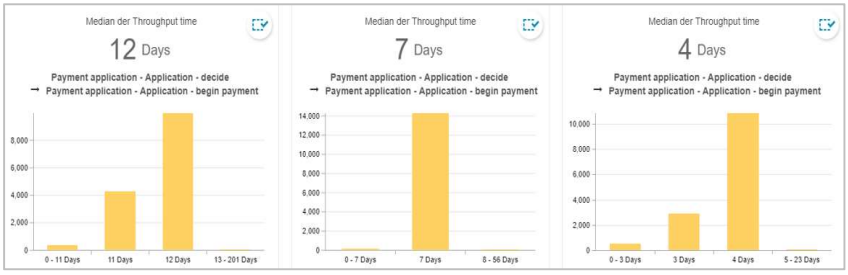
The decision-making process is seen to be the fastest in 2015 and takes most time in 2017 (refer Fig. 23 (a)). The main contributors in this delay in 2017 are the document types 'Geo parcel' and 'Payment application' (refer Fig.23 (b)).



(a) (b)

Fig. 23. (a) Distribution of applications by throughput time from start to decide across years (b) Mix of throughput time across document types for applications from start to decide across years

Additionally, we see that the time between decision ('Payment application-Application-decide') to the beginning of payments ('Payment application-Application-begin payment') is least in 2017 and most in 2015 (refer Fig. 24). This value has decreased 67% in two years. Hence, the delay in taking a decision in 2017 is partially compensated by taking lesser time from the decision to the payment step.



(a) 2015 **(b) 2016** **(c) 2017**

Fig. 24. Distribution of applications by throughput time from decide to payment across years

Conclusion

We have shown the existence of concept drift across years using ProM, i.e., there indeed are changes in process flow across years. We further identified what these changes exactly are in terms of control flow, and probable reasons for the same. The impact that these differences have on the process has also been quantified using important metrics. We see that as we move ahead in timeline, effort and time taken to reach the decision stage in the process increases. But at the same time, the resources utilized for the same decreases.

Future Scope of Work

Since we limited ourselves to identification of 'Sudden' concept drift in terms of 'Control flow' only, much more can be done to analyze the process from various other perspectives and other types of concept drifts in the future. Further, the analysis can be carried out at a more granular activity level, concatenation of three fields as mentioned in section 3.1. This would give us more insight as to where the process exactly drifts across years.

7 References

1. European Commission Direct Payments for Farmers 2015-2020, https://ec.europa.eu/agriculture/sites/agriculture/files/direct-support/direct-payments/docs/direct-payments-schemes_en.pdf
2. European Commission Direct Payments Homepage, https://ec.europa.eu/agriculture/direct-support/direct-payments_en
3. BPI Challenge 2018 Homepage, http://www.win.tue.nl/bpi/doku.php?id=2018:challenge
4. Data Experts Homepage, https://www.data-experts.de/
5. Basic Payment Scheme, https://ec.europa.eu/agriculture/glossary/basic-payment-scheme_en
6. BPI Challenge 2018 Forum Discussion, http://www.win.tue.nl/promforum/discussion/986/question-on-inspection
7. Process mining tool Myinvenio homepage, https://www.my-invenio.com/
8. R.P. Jagadeesh Chandra Bose, Wil M.P. van der Aalst, Indre Zliobaite and Mykola Pechenizkiy : Handling Concept Drift in Process Mining, wwwis.win.tue.nl/~publications/p623.pdf
9. BPI Challenge 2018 Forum Discussion, http://www.win.tue.nl/promforum/discussion/1017/i-have-two-questions-regarding-severe-penalties-and-randomly-selected-applications#latest
10. Variance Inflation Factor, https://en.m.wikipedia.org/wiki/Variance_inflation_factor
11. Receiver Operating Characteristic Curve, https://en.m.wikipedia.org/wiki/Receiver_operating_characteristic