Use of statistical methods to investigate the database of foreign currency market transactions

Amir Khatib and Arel Mazouz *[1]

Summary

Working with Big Data databases, the enormous volume of which makes it difficult to use common data analysis methods, requires the use of advanced statistical tools to perform data quality control and draw insights from information.

The Bank of Israel's Information and Statistics Department manages a database of Big Data on derivative financial instruments in the forex and interest rate markets, which contains information on transactions in these derivatives on the OTC market. The database is used primarily to gain knowledge of the market and to develop products that support the decisions of the Bank's policy makers.

This study reviews several tools and methods that the Department uses to optimize, process, and generate insights from the information existing in the derivatives database. The study also describes several data science models that are used to identify irregular reports, complete missing data, classify market participants, and enable the Department to draw insights that inform policy making.

* Information and Statistics Division of the Bank of Israel.

[1] The Department's Statistical Methods and Data Science Unit was also involved in the process of deliberating on and implementing the models and methods described in this study.

1. Introduction

The Bank of Israel's Information and Statistics Department ("the Department") manages an itemized database on derivate financial instruments in the foreign exchange and interest rate markets ("the Derivatives Database")[2]. The Department receives information on transactions in foreign exchange and interest rate derivatives on the OTC market from financial intermediaries in and outside Israel. These are intraday trading data that are received on a daily basis. The financial intermediaries that are subject to a reporting requirement ("the Reporting Entities") are Israeli banking corporations and financial intermediaries whose operations in the shekel-based foreign currency market exceed US $15 million per day[3].

The information entered into the database may reach several million records per year and creates a Big Data database due to its historical depth and the extensive information available on each transaction. Advanced statistical tools are therefore necessary to control the quality of the data and derive meaning from them. The aim of this study is to review several tools that the Department uses for this purpose. To draw a comprehensive picture that reflects the events on the OTC market and enables insights to be drawn from the reported data, the Department adopted statistical tools that assist in classifying market participants based on specific features that were determined on the basis of the models described below.

The first section of this study includes a brief overview of the data entered into the database, in order for the reader to gain a general familiarity with the system. The second section presents several examples of insights that were produced from the database using statistical tools. For example, we will see that a specific pattern of activity characterizes companies that primarily import of goods and services, and this pattern is significantly different from the typical pattern of activity of companies that export goods and services.

The third section concerns the manner in which insights are generated from the data. The system was designed to provide the Bank of Israel with information on the foreign exchange market that is as detailed as possible, and to yield insights that would assist the Bank in making optimal decisions concerning this market. This is the ultimate aim of the Derivatives Database: to give the Bank tools in order to perform its functions.

The fourth section briefly presents several statistical controls that were assimilated into the Derivative Database in order to identify irregular data. This section discusses the problems confronting the Department, and describes the statistical models that were tested and the model that was ultimately selected for use. Although this is not the topic of this study, we believe that a brief review of these controls is important because all statistical work begins with an examination of the quality of the data.

2. Background to the data system

The Derivatives Database monitors shekel-based trading on the OTC market. The database includes several main dimensions of these trades—financial instruments, underlying assets, time, and sectors—that help the Bank of Israel specify the nature of activities on the shekel/forex market and develop its policy accordingly.

The transaction data in the system come from three types of information sources:

- Domestic banking corporations, which have reported their transactions since 20008;
- Foreign financial institutions, which have reported their transactions since 2017;
- Domestic financial intermediaries, which have reported their transactions since 2017.

Data are received daily and include details of all the transactions made in the preceding business day. Because reports are received from different time zones, it was determined for the sake of uniformity that transactions are reported on the basis of Universal Time (UTC). Therefore the daily reports include transactions executed between 00:00 and 23:59 UTC on the day of business (T), and the required information is received on the following business day (T+1).

[2] For additional information, see Statistical Bulletin 2018.

[3] For more information, see the Bank of Israel Order.

The system includes about 40 reporting entities and tens of thousands of active OTC market participants. As noted above, the system receives several million records each year.

The new reporting format was adjusted to the international standard developed by the ISDA[4], the organization that leads the promotion of procedures for standardization of OTC trading. The new format contains about 40 fields related to three factors:

1. The reporting entity
2. The participant or customer
3. The transaction – trading time (second level), instrument, contract expiration date, transaction price, nominal amount, and transaction value.

The Bank of Israel generates many products based on the data in the Derivatives Database, some of which are published in the periodic notices issued by the Bank on its activities in the foreign exchange market. Examples of products generated using the system include estimates of aggregate foreign exchange purchases of primary sectors, daily trading volume in the foreign exchange market by instrument type, and standard deviation of changes in shekel-dollar options.

3. Statistical models as a tool for classification

The transaction data are entered into the system after undergoing all of the necessary logical controls. At this stage they may be subject to various analyses. Because the database contains several million records, appropriate Data Science methods and algorithms must be used. These are mainly divided into two types:

Supervised learning – These methods are appropriate for a data model with existing tables or classifications, where the aim is to classify new observations on the basis of existing classifications of previous observations.

Unsupervised learning – In this case no prior classification exists, and the aim is to divide the data into clusters based on the internal structure of the data and the common patterns that the algorithm is able to identify.

The advantage of unsupervised learning is in its lack of any fixed preconception about the results, and the researcher who uses this type of method approaches the data with an open mind toward various types of results. Still, it is possible that no division emerges that is consistent with the data's underlying logic (see below for additional information). In contrast, supervised learning is more "rigid" in that the desired outcome is defined in advance, yet the investigator may "assist" the model by assigning varying weights to the variables to improve the model's explanatory power.

To illustrate the methods, the example that accompanies this section is an attempt to classify the various business sector participants that executed transactions in the foreign exchange market into different groups.

First, the business sector is divided into several clusters using an unsupervised learning method known as Principal Component Analysis (PCA) together with k-MEANS clustering.

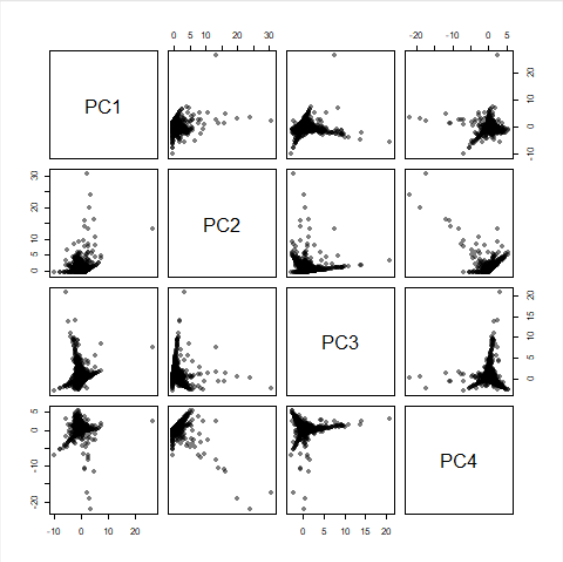
[4] The ISDA created a uniform contract for derivatives trading. For additional information see https://www.isda.org/#

A. Cluster analysis using unsupervised learning

In the effort to classify the participants into groups based on the features of their transactions, we used PCA, which generates a graphic depiction of the observations in the database by reducing the number of variables displayed in the graph. This method takes all the principal variables (defined by the user, e.g., trading volume, forex purchases, etc.) and generates new variables such that each new variable is a linear combination of the principal variables. In this way the method creates a reduced set of variables that contain the majority of the variance that exists in the selected dataset.

To distinguish between the various groups in the business sector using PCA, we entered 8 principal transaction features, including average daily trading volume, average daily net forex purchases, average/median range of transactions, etc. Our method reduced these into four new variables (that are linear combinations of the original variables) that cover approximately 87 percent of the existing variance. Figure 1 below shows the various combinations of the 4 variables on a two-dimensional grid. Each graph presents a division of the observations in the dataset based on two explanatory variables. For example, the top right graph shows the division of the dataset based on PC1 and PC4. These graphs make it simple to distinguish between the various clusters in the business sector (if any exist) based on the raw data. It is important to note that a division into clusters using PCA is based on statistical variance, and therefore, this division will not necessarily correspond to an economic division of the dataset.

Figure 1: Grid of the Principal Components (PCA)



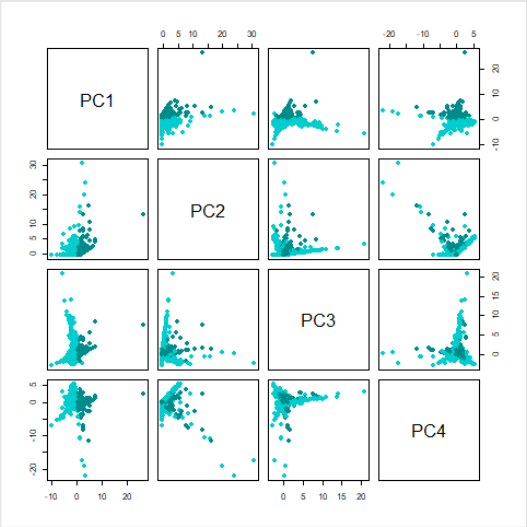
SOURCE: Based on reports by the banks

Although it is possible to divide the participants in the two-dimensional spaces into several groups, we see that there is no clear distinction between the groups in any classification scheme. For example, the scatter plot for PC3 and PC4 shows three “arms” that emanate from the central core of the distribution into different directions, such that we may state that they divide the data but this division is not statistically significant. We therefore turn to k-MEANS, another unsupervised learning method.

K-MEANS is a method that is commonly used in data science for clustering. The aim of this method is to partition the data based on weighted cluster-centers, where the user determines the number of cluster-centers, and each cluster-center represents clusters of data. By correctly selecting the number of cluster-centers, it is possible to identify various groups within the data set.

To divide the participants in the business sector into groups, a different number of clusters was selected for each analysis (the method was applied for 2, 3, and 4 clusters). For every predetermined number of cluster-centers, the k-MEANS method divided the data set into groups (see Figure 2, which describes a division based on two cluster-centers), but there was no economic rationale to explain these divisions, for example, between importers and exporters. Therefore we now turn to describe a classification process using Linear Discriminant Analysis (LDA), a method of supervised learning.

Figure 2: Grid of the Principal Components Grouped According To the K-Means method



SOURCE: Based on reports by the banks

B. Classification into groups using supervised learning

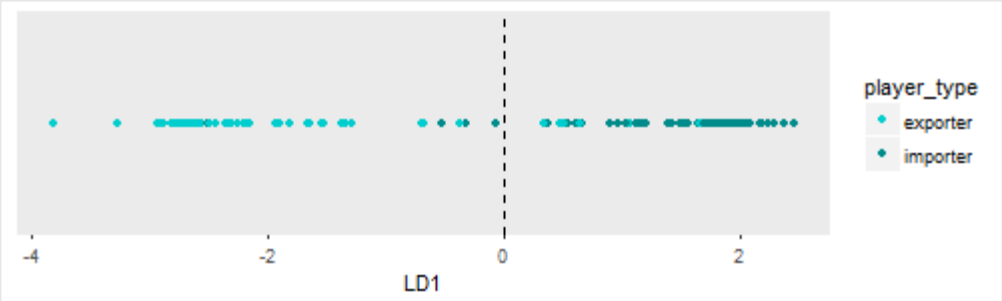
When a division into groups is known in advance, even if it is based on a sample of the data, we can use LDA to identify the properties that distinguish one group from another. LDA calculates a linear combination of the explanatory variables that creates the best distinction between the predefined groups (the target variable). We can use this linear combination to classify data that were not included in the initial sample, as well as new data.

In the example of the classification of business sector participants into groups that accompanies this section, we might have arrived at an initial division of companies engaged in imports versus companies engaged in exports. Using other Bank of Israel databases as well as specific economic properties, we constructed a list of companies, distinguishing between import companies and export companies. For example, we would expect a company engaged primarily in import to purchase more foreign currency than it sells, compared to a company engaged primarily in export, which we would expect to sell more foreign currency than it purchases. Accordingly, export companies were defined as companies that export more goods than they import, while import companies were defined as the opposite case, companies that import more goods than they export.

We then used the same explanatory variables that characterize each participant, such as annual trading volume, number of operating days as a share of total business days, net purchases, use of the dollar as a share of all currencies used, average range of transactions, etc. An algorithm was applied to all these variables in order to identify the greatest separation between the two predetermined groups. The result was a linear combination of explanatory variables that best distinguishes between the two groups.

Using that one-dimensional combination, we can see (Figure 3) a clear distinction between the nature of activities of importers and exporters. This allows us to classify new unclassified data automatically and reliably on the basis of their characteristic market activities.

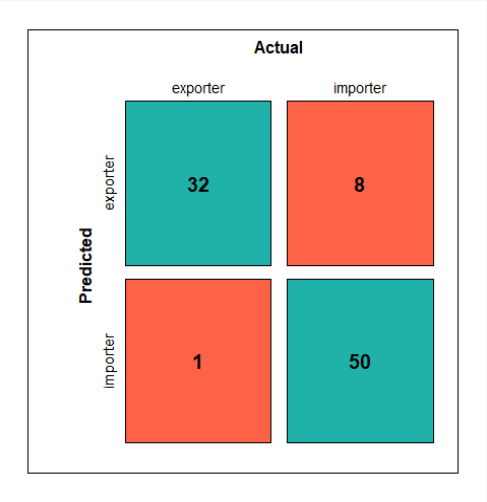
Figure 3: Importer/Exporter Grouping by Linear Combination Using LDA



SOURCE: Based on reports by the banks

In view of the satisfactory results of the model, we applied this linear combination to a predefined test group: companies in the business sector with an importer or exporter classification (this classification was based on other databases and is similar to the sample group) and that were not part of the sample used to calculate the linear combination. The model classified these companies according to the coefficients for each variable that were calculated in the LDA. We compared the classification based on the model to the predetermined classification, and the findings show (see Figure 4) that the model generated a good prediction of the division of companies into import and export firms on the basis of their market activities. Therefore, a decision was made to use this linear combination on a regular basis to classify new business sector participants that had no predetermined classification.

Figure 4: Confusion Matrix



SOURCE: **Based on reports by the banks**

4. Generating insights from the data

Until now this study focused on classifying and improving the data, but no matter how high the quality of a database, it is meaningless if it does not yield insights. Therefore, to generate insights and products from a Big Data system such as the Derivatives Database, we integrated additional statistical tests into the system.

A. Identifying differences between sectors

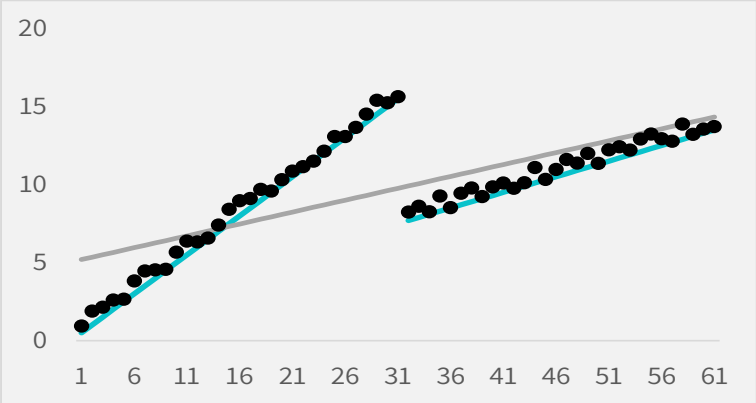
After classifying the data into groups using LDA, which was described in the previous section, we examined the coefficients of the linear combination that the model selected to classify the participants. This analysis found that the groups differ on several factors, with the primary factor being that exporters are active in the market on many more days than importers, but importers are more active than exporters on the days with irregular changes in foreign exchange rates and also enter into more long-term transactions. The groups also differ in their net foreign exchange purchases: In line with the nature of their business activities, exporters sell foreign currency while importers purchase it.

B. Identifying structural breaks in time series data

One of the most significant tools that may support policy making is the ability to distinguish between various time periods defined on the basis of predetermined properties (such as the ability to distinguish between periods of recession and growth). Similarly, by breaking down a specific sector's activities into various time periods we may learn a great deal about the nature of that sector's activities. To illustrate, if we could determine that the business sector purchases more foreign currency when the shekel is appreciated (due to importers' need to protect themselves against an additional appreciation), we would be able to better assess how an appreciation or depreciation affects the activities of that sector.

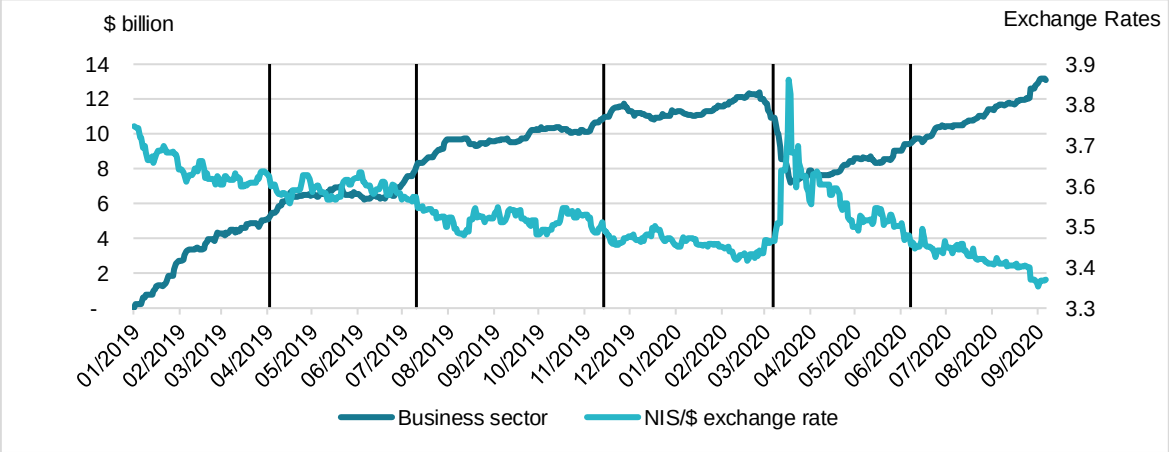
To identify various trends, we integrated a statistical method known as a structural break. The method is based on an algorithm, developed by Bai and Perron (2003) that calculates several regression models to obtain the best division of a time series. The division is determined endogenously and leads to a model that minimizes the sum of squared residuals and thus better explains the time series. For example, we can observe in Figure 5 that in this case two regression lines explain the time series better than the regression line that represents the entire period.

Figure 5: Graphic Illusration of the Bai & Perron Method



This method was applied to time series of aggregate foreign exchange purchases by the business sector. The findings follow:

Figure 6: The Representative NIS/$ Exchange Rate and Aggregate Net Foreign Exchange Purchases by the Business Sector, January 2019–September 2020



SOURCE: Bank of Israel

We obtained a statistically significant division for each group and identified structural breaks that divide the period into several subperiods. Furthermore, the structural breaks are identical in both the aggregate purchases of the real sector and in the shekel/dollar exchange rate data, which means that the trends in both time series are characterized by the same six subperiods, which apparently is an indication of reciprocal effects between the business sector's activities and changes in the exchange rate.

C. Identifying irregular market activity

To identify irregular activity of a specific sector or participant, we constructed a warning system that takes into account the primary trading features of each group or participant (trading volume, net foreign exchange purchases, etc.). A test based on the MAD method is performed on a daily basis to identify unusual activity. If unusual activity is identified, the system verifies that the data represent a true irregularity in the participant's or sector's activity and not a reporting error. Then, the irregular activity is analyzed according to general variables related to the foreign exchange market and that potentially affect the activities of sectors or participants on that day. This analysis makes it possible to find "the story" behind the irregular activities in order to better understand the market and its dynamics.

For example, Participant A is engaged in export activities that generally feature a moderate sale of dollars over the entire year (for instance, US $10 million at a time), probably based on his revenues from overseas. If Participant A sells a large amount of dollars on a certain day (for instance, US $100 million), the system will generate an alert that Participant A was involved in irregular activity, together with the main features of his activities (mean, median, standard deviation, etc.). The data are then inspected manually and Participant A is contacted if necessary. Using this method, it is possible to identify days of irregular activity or changes in participant' patterns of behavior, which makes a significant contribution to the Bank's understanding and monitoring of the market.

5. Data optimization

We believe it is important to add to this study a brief review of several of the system's logical controls. The significance of this review stems from the fact that any statistical model, however good, is useless if the database on which it is tested contains missing or incorrect information.

When the daily data reports are received from the reporting entities, several logical controls are performed to verify that all mandatory fields are complete, the data match the required format of each field, and the transaction meets several basic logical conditions (e.g., the transaction execution date cannot be later than the payment date). Furthermore, after the data are entered into the system, they are cross-checked with the reports from other reporting entities (for example, a transaction reported between two reporting entities must be reported by both using the same information, such as date, amounts, exchange rate, etc.). This test makes it possible to identify missing or incorrect reports and therefore assists in improving the quality of the data in the system.

However, as described below, data that pass these controls are not necessarily error-free. We therefore briefly describe several tools used to identity irregularities in the data.

A. Identifying irregular data

Transaction amount – As noted above, an incorrect transaction may nonetheless satisfy all the conditions defined in the controls and be entered into the system. For example, a US $1 billion conversion transaction is possible but it if is reported for an entity whose annual volume of transactions totals several tens of millions of dollars, the transaction may be suspected as being incorrect.

After an extensive inspection of the data and application of multiple statistical methods, it was found that the distribution of reported transaction amounts was close to normal after a log transformation. Furthermore, it appears that the use of averages and standard deviations leads to a larger standard deviation due to irregular observations, which makes it difficult to determine the number of standard deviations that defines a datum as irregular. Therefore, we used MAD to identify irregular transaction amounts, as this method is not affected by irregular observations.

Every day a test of irregular transactions is performed, and these are compared to various features of the market, the participant, and other macroeconomic factors in order to determine whether the transaction is a logical one or a reporting error. If it appears to be a reporting error, the reporting entity is contacted to confirm or correct the reported transaction.

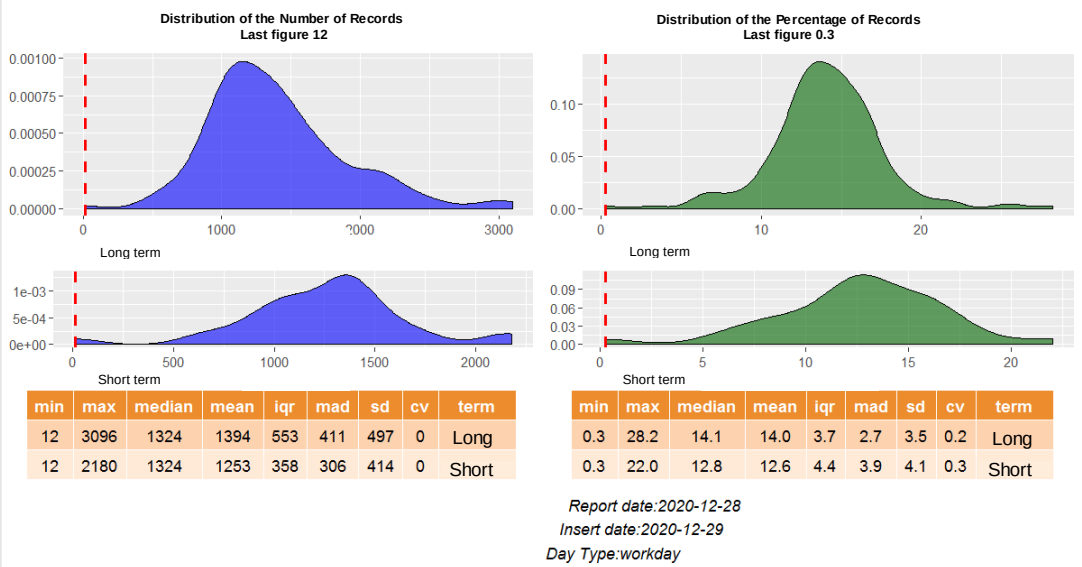
Number of reported records – As noted above, each reporting entity transmits a daily report of the transactions it executed on the OTC market in the preceding business day. Occasionally a reporting entity will be affected by a malfunction that causes only part of the transactions to be reported in the daily report file. In order to be able to identify such reports (which are more difficult to identity than nonreporting), the Department was required to resolve several issues, such as how to distinguish between heavy and light trading days, how to account for exogenous changes that affect shekel trading, how to determine the benchmark index, and others.

To select the model with the best fit, two approaches were tested over time – a test of the quantity of transactions reported by a reporting entity compared to the number of transactions reported by that entity in the short term and in the long term, and the proportion of transactions reported by the reporting entity out of the total number of reported transactions, distinguishing between domestic and foreign reporting entities (this approach assumes a connection between the change in the number of transactions reported by one reporting entity and the change in the number of transactions reported by all reporting entities).

As Figure 7 shows, the distribution of the reporting entities' reports has a long right tail (known as skewness). This is not a surprising finding, as the number of reported transactions is necessarily non-negative and generally will be centered around the mean/median with a large number of observations that comprise the right tail. However, due to this distribution structure, the mean and standard deviation cannot be used to identify irregular data.

We applied MAD to the data, where an irregular report was defined as a deviation of MAD × 2 from the median. As indicated in Figure 7, the red line is the most recent report by the reporting entity, and it is preceded by the distribution over time (short term above long term). The right panel shows the distribution of the reporting entity's transactions as a percentage of the total number of reported transactions, and the left panel shows the distribution of the number of transactions reported by the reporting entity. Furthermore, each analysis is accompanied by a table containing the main data (mean, median, minimum, maximum, etc.).

Figure 7: Irregular Data in the Reports – Distribution Of the Number of Records and their Relative Proportion



min	max	median	mean	iqr	mad	sd	cv	term
12	3096	1324	1394	553	411	497	0	Long
12	2180	1324	1253	358	306	414	0	Short

min	max	median	mean	iqr	mad	sd	cv	term
0.3	28.2	14.1	14.0	3.7	2.7	3.5	0.2	Long
0.3	22.0	12.8	12.6	4.4	3.9	4.1	0.3	Short

After some time in which both approaches were tested, findings showed that a reporting entity's percentage of the total number of market transactions showed greater stability, and this approach was therefore selected. This process proved to be effective. Since this process was implemented, a large number of irregular reports were identified. Without this process these irregular reports would not have been discovered and would have adversely affected the quality of the data in the system. For example, in June 2020, an irregular report was discovered, and an inquiry made with the reporting entity revealed that only 3% of the executed transactions had been transmitted in that report.

6. Summary

This study presents some of the tools and methods used by the Bank of Israel's Information and Statistics Department to optimize, process, and generate insights from the information in the unique Derivatives Database system. The quantity of information that is collected calls for the use of advanced statistical tools to control and derive meaning from the data that are received.

The first part of this study described several models that are used to complete the data and generate insights from it in order to obtain a complete and comprehensive picture of the activities of the various market participants. The methods described come from the world of data science and include methods such as PCA in combination with k-MEANS to identify clusters of participants, and LDA to classify various business sector participants. We used the LDA method to identify the properties of subgroups within the business sector, and found statistically significant differences between importers and exporters.

In the second part of this study we reviewed several statistical controls that were integrated into the system to identity irregular data, and also reviewed the concerns that guided the development of the controls and the selection of the most appropriate statistical tool in each case.
The aim of this study was to provide an overview of several of the statistical methods used by the Department to control and process data.