Grant Agreement Number: 825225

Safe-DEED

www.safe-deed.eu

D7.9 Deliver Qualified Synthetic Data (V2)

Deliverable number	*D7.9*
Dissemination level	*Public*
Delivery date	*30 November 2020*
Status	*Final*
Author(s)	*Tobias Welling, Alexander Georg*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825225.

Changes Summary

Date	Author	Summary	Version
21.11.2019	Alexander Georg	Finalised version after review	1.2
28.10.2020	Tobias Welling	Second Draft (updated info on data collection and Conclusion) as marked in light grey	2.1
13.11.2020	Patrick Ofner	Review	2.2
26.11.2020	Tobias Welling	Final Version	2.3

Executive summary

Synthetic data makes an important contribution to data analysis and research. Qualified Synthetic Data is a meaningful representation of real data and anonymized for analysis purposes.

The source data is collected internally at Infineon from an order tracking tool that measures order lead times. In this context, all relevant parameters for Lead Time Based Pricing (LTBP) are synthesized. These parameters include Order Entry date, Requested Order Lead Time, Confirmed Order Lead Time and Order Volume.

The composition of the data and the procedure for the creation are described in this deliverable. This deliverable is based on prior submitted deliverable 7.3 and thus features the same structure. It has been updated in certain aspects. The target to offer QSD for several product lines and different ranges of business months could not be completed, since the amount of needed and then generated qualified synthetic data would have exceeded the usual data handling capabilities. Instead, it has been decided to provide an updated QSD dataset which includes the full data picture for the same product line from beginning of 2019 until up until the publication of this deliverable. To put the data sizes into perspective, one excel file including only the QSD is already bigger than 36 Mega Bytes and only include the order data for one product line. Thus this deliverable focusses more on the data retrieval algorithm.

Table of Contents

List of Figures

List of Abbreviations

Business-month	BM
Data Management Plan	DMP
Infineon	IFX
Lead Time based Pricing	LTBP
Product line	PL
Qualified Synthetic Data	QSD
Work Package	WP

1 Introduction

Customer- and product specific order data serve as the basic input data for a possible introduction of Lead Time based Pricing (LTBP). Relevant parameters thereby include relevant dates such as Order Entry Date and Customer Wish data, as well as order volumes and contractual agreements for the lead time. In order to share the data with project partners without giving out sensible customer or product related information and to provide meaningful data for testing algorithms related to Lead Time based Pricing, qualified synthetic data has been created for the order data.

Qualified Synthetic data (QSD) are commonly described as "any production data applicable to a given situation that are not obtained by direct measurement" (Parker, 1984). QSD provide an intelligent way to create test data that reflect more than just the structure of the dataset, but also preserve the semantics and meaning of the data while guaranteeing privacy (Machanavajjhala et al., 2008). Test data primarily means input data for algorithms as the pricing algorithm for Lead Time based Pricing in our case, but could potentially also serve training purpose in machine learning applications. Besides methods such as simulation models and neural networks, statistical methods represent a key way to create synthetic data. Infineon has chosen to apply the statistical method for the creation of synthetic data. Thereby the key element is to design a statistical model from the original data and then create anonymized samples from the model (Machanavajjhala et.al. 2008).

Chapter 2 describes in detail the collected data and the way of creating the synthetic data. Chapter 3 explains the location of the stored data and the means of data security. Chapter 4 elaborates on the data usage.

2 Data composition and creation

This chapter provides some background information and a short description of the collected data that is distributed using the Qualified Synthetic Dataset. The second part elaborates on the statistical procedure for creating qualified synthetic data from the real-world order data.

2.1 Data Collection

The data is generated by an algorithm which measures customer Order Lead Times. Generally companies use enterprise resource planning systems like SAP[1] for tracking and managing company data, including order data. Order data in the semiconductor industry that Infineon is operating in, is characterized by long lead times and frequent order changes by customers until the final changes. The SAP system is not sufficiently capable of dealing with these frequent changes in e.g. volumes or Requested Delivery dates. When new items are added to an existing order, order entry dates are not updated and receive the entry date of the initial order. Therefore an algorithm is necessary to calculate and measure the correct Order Lead Times.

In order to solve the challenges of adequately measuring lead time, an algorithm has been created to properly assess the different lead times as they are key input variables for revenue management and an enhanced pricing algorithm. As changes of orders need to be tracked on a daily basis to generate proper results, the report of pipeline of open orders ("Auftragsbestand") was downloaded every day from SAP database and the newest report was compared to the previous one whether some parameters were modified. More elaborately explained, the data collection starts with filtering the backlog report for the latest entries of orders, namely the ones entered on the day of the report creation date, thus, only representing orders of a single day. Thereafter, on the following days, the filter was adapted to show a certain time frame, starting from the first day of the analysis up to the newest orders added to the report.

[1] https://www.sap.com/corporate/de.html

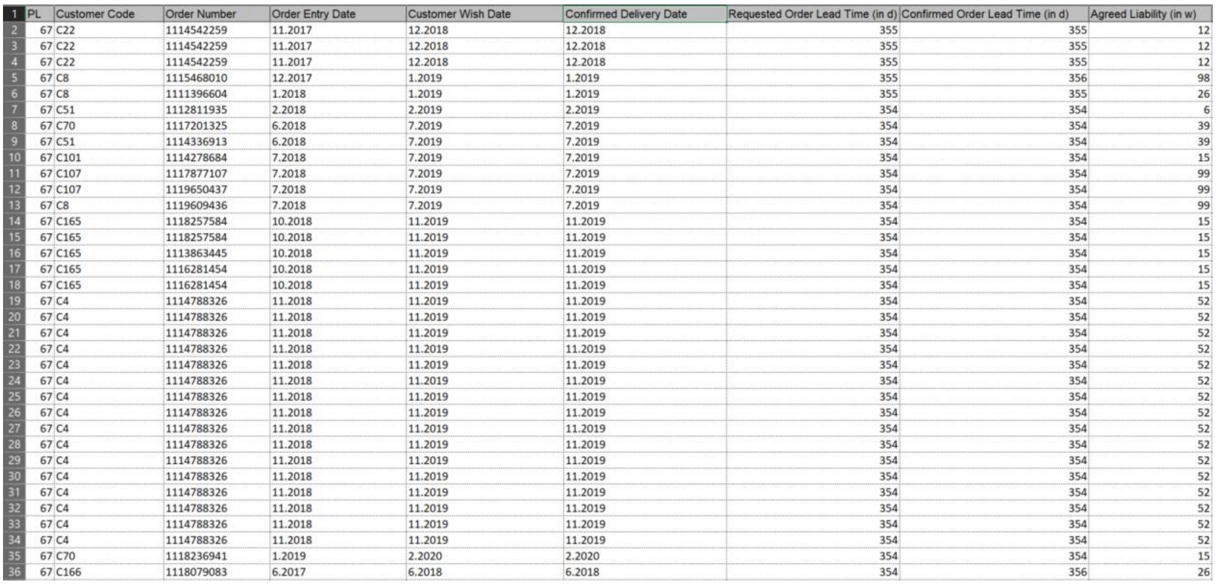
In consequence, the data collection has been undertaken on an enrolling basis: New orders are added to the master data set, whereas already existing orders, which were previously entered, are inspected whether any changes occurred in the latest report. The analysis starts with orders of a single day as it cannot be tracked in retrospect whether e.g. amount changes have been made between the order entry date and the report creation date which would once again lead to an incorrect lead time measurement.

The algorithm was automated, since its manual execution for each order would take a tremendous amount of time. The automation code is based on the programming language Visual Basics (VBA) in Microsoft Office Excel in several combined macroinstructions (macros).

As Lead Time data is confidential, Qualified Synthetic Data was be created that share the same distribution and characteristics, but do not depict sensible customer information. The distribution of products and customers will be taken into account as well, but in encoded form both for security and confidentiality reasons. The final table of data will include Product Line, Business Month, Order Entry Date, Requested and confirmed Order Lead Time, customer name encoded, product name encoded, Order number and Order volume. The data is stored in Excel format. The Excel format also enables sharing and long-term access. The size has been initially be approximately 5.5 MB of data, including monthly data for the year-to-date and the previous business year, but fluctuates based on the published version on QSD, since the more recent version includes data from a larger time frame. Data for one Product line is derived and converted for the other Product lines as the distribution is almost identical. Other third-party data has not been used.

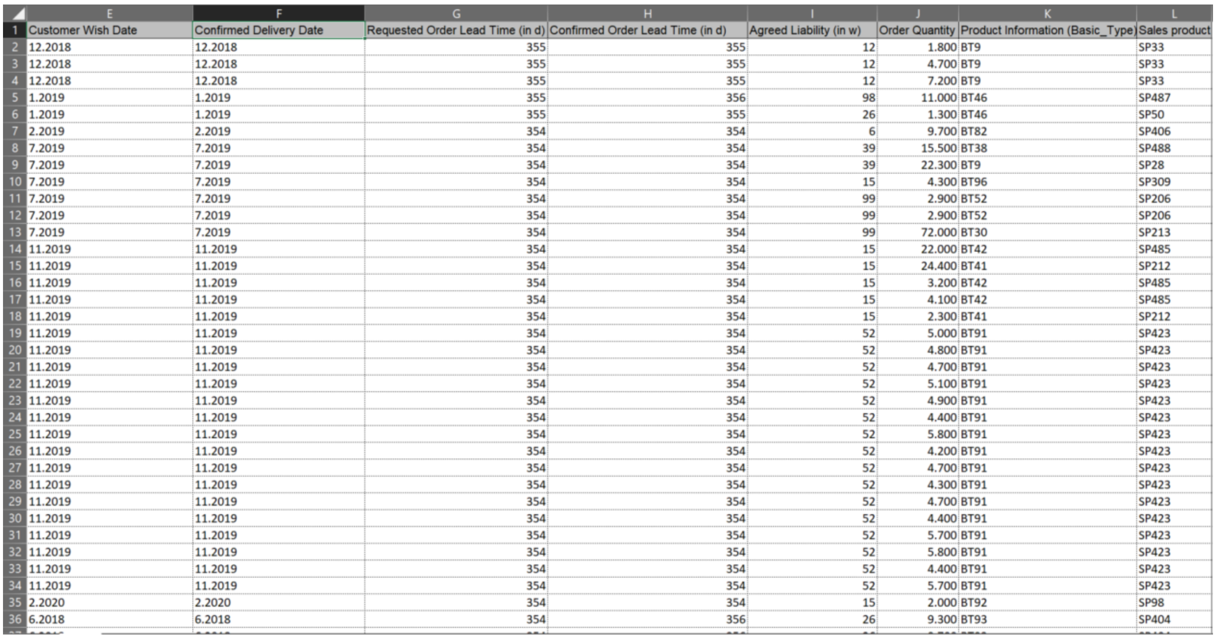
2.2 Data Creation

A statistical method is used to generate the synthetic data. Thereby data points of the measured Order Lead Times are ordered to identify distribution functions that share the main characteristics with the real data. The procedure will be described using an example. The synthetic data is generated column-wise. Requested and Confirmed Order Lead Time are separately ordered in descending order. Extreme outliers are excluded using a 0.99 confidence interval. As an example, this results in an exclusion of values larger than $\mu+2.6*\sigma$ (for 40,000 data points), with μ being the mean of the column and σ being its standard deviation. In a next step, the data is fitted to distribution functions and in order to achieve a high goodness of fit, only distribution functions with a coefficient of determination of $R^2 > 0.98$ are acceptable. The synthetic data is then created by sampling from the identified distribution functions. Customer Wish Dates are computed by adding the synthetic Requested Order Lead Time on top of the order entry date. Confirmed dates result from adding the Confirmed Order Lead Time to the order entry date. Customers and products are encoded by assigning integer number codes to the names. Data quality and validity is insured by expert peer review after each iteration step when generating the synthetic data and after the use of R^2 for good fit between real and synthetic data. Figures 1 and 2 show a screenshot of the data table of version one that is uploaded on Zenodo.



1	PL	Customer Code	Order Number	Order Entry Date	Customer Wish Date	Confirmed Delivery Date	Requested Order Lead Time (in d)	Confirmed Order Lead Time (in d)	Agreed Liability (in w)
2	67	C22	1114542259	11.2017	12.2018	12.2018	355	355	12
3	67	C22	1114542259	11.2017	12.2018	12.2018	355	355	12
4	67	C22	1114542259	11.2017	12.2018	12.2018	355	355	12
5	67	C8	1115468010	12.2017	1.2019	1.2019	355	356	98
6	67	C8	1111396604	1.2018	1.2019	1.2019	355	355	26
7	67	C51	1112811935	2.2018	2.2019	2.2019	354	354	6
8	67	C70	1117201325	6.2018	7.2019	7.2019	354	354	39
9	67	C51	1114336913	6.2018	7.2019	7.2019	354	354	39
10	67	C101	1114278684	7.2018	7.2019	7.2019	354	354	15
11	67	C107	1117877107	7.2018	7.2019	7.2019	354	354	99
12	67	C107	1119650437	7.2018	7.2019	7.2019	354	354	99
13	67	C8	1119609436	7.2018	7.2019	7.2019	354	354	99
14	67	C165	1118257584	10.2018	11.2019	11.2019	354	354	15
15	67	C165	1118257584	10.2018	11.2019	11.2019	354	354	15
16	67	C165	1113863445	10.2018	11.2019	11.2019	354	354	15
17	67	C165	1116281454	10.2018	11.2019	11.2019	354	354	15
18	67	C165	1116281454	10.2018	11.2019	11.2019	354	354	15
19	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
20	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
21	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
22	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
23	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
24	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
25	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
26	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
27	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
28	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
29	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
30	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
31	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
32	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
33	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
34	67	C4	1114788326	11.2018	11.2019	11.2019	354	354	52
35	67	C70	1118236941	1.2019	2.2020	2.2020	354	354	15
36	67	C166	1118079083	6.2017	6.2018	6.2018	354	356	26

Figure 1: First part of the uploaded table of synthetic data



	E	F	G	H	I	J	K	L
1	Customer Wish Date	Confirmed Delivery Date	Requested Order Lead Time (in d)	Confirmed Order Lead Time (in d)	Agreed Liability (in w)	Order Quantity	Product Information (Basic_Type)	Sales product
2	12.2018	12.2018	355	355	12	1.800	BT9	SP33
3	12.2018	12.2018	355	355	12	4.700	BT9	SP33
4	12.2018	12.2018	355	355	12	7.200	BT9	SP33
5	1.2019	1.2019	355	356	98	11.000	BT46	SP487
6	1.2019	1.2019	355	355	26	1.300	BT46	SP50
7	2.2019	2.2019	354	354	6	9.700	BT82	SP406
8	7.2019	7.2019	354	354	39	15.500	BT38	SP488
9	7.2019	7.2019	354	354	39	22.300	BT9	SP28
10	7.2019	7.2019	354	354	15	4.300	BT96	SP309
11	7.2019	7.2019	354	354	99	2.900	BT52	SP206
12	7.2019	7.2019	354	354	99	2.900	BT52	SP206
13	7.2019	7.2019	354	354	99	72.000	BT30	SP213
14	11.2019	11.2019	354	354	15	22.000	BT42	SP485
15	11.2019	11.2019	354	354	15	24.400	BT41	SP212
16	11.2019	11.2019	354	354	15	3.200	BT42	SP485
17	11.2019	11.2019	354	354	15	4.100	BT42	SP485
18	11.2019	11.2019	354	354	15	2.300	BT41	SP212
19	11.2019	11.2019	354	354	52	5.000	BT91	SP423
20	11.2019	11.2019	354	354	52	4.800	BT91	SP423
21	11.2019	11.2019	354	354	52	4.700	BT91	SP423
22	11.2019	11.2019	354	354	52	5.100	BT91	SP423
23	11.2019	11.2019	354	354	52	4.900	BT91	SP423
24	11.2019	11.2019	354	354	52	4.400	BT91	SP423
25	11.2019	11.2019	354	354	52	5.800	BT91	SP423
26	11.2019	11.2019	354	354	52	4.200	BT91	SP423
27	11.2019	11.2019	354	354	52	4.700	BT91	SP423
28	11.2019	11.2019	354	354	52	4.300	BT91	SP423
29	11.2019	11.2019	354	354	52	4.700	BT91	SP423
30	11.2019	11.2019	354	354	52	4.400	BT91	SP423
31	11.2019	11.2019	354	354	52	5.700	BT91	SP423
32	11.2019	11.2019	354	354	52	5.800	BT91	SP423
33	11.2019	11.2019	354	354	52	4.400	BT91	SP423
34	11.2019	11.2019	354	354	52	5.700	BT91	SP423
35	2.2020	2.2020	354	354	15	2.000	BT92	SP98
36	6.2018	6.2018	354	356	26	9.300	BT93	SP404

Figure 2: Second part of the uploaded table of synthetic data

3 Data storage and data security

In this chapter, data storage of the Synthetic Data and responsibilities for creation and maintenance will be described. Moreover, the means of data security are described in the second part. .

3.1 Data storage

The data is uploaded on Zenodo, an open repository. An overview of findable DOIs is given in the following: The first version of QSD has been uploaded on the 8th November 2019 and can be found under https://doi.org/10.5281/zenodo.3532890 or the DOI: 10.5281/zenodo.3532890. The second version of QSD has been published on the 10th November 2020 and can be subsequently found under https://doi.org/10.5281/zenodo.4032980 and the DOI: 10.5281/zenodo.4032980. The second version contains up to the publication of this deliverable the newest data and is as such larger in size than the version 1 of QSD.

The participants from WP7 (Infineon Technologies AG) of the Safe-DEED project are responsible for implementation, review and revision of the Data Management Plan (DMP). Each management step is thereby fully accountable to WP7.

3.2 Data security

Infineon follows a thorough set of rules and regulations in order to protect the data of customers, employees and involved third parties. All required data privacy and security laws set by both German and EU law, are followed. Infineon has made all necessary changes and established protocols to ensure its compliance with the new GDPR requirements and recommendations. All processes are governed and undertaken by Infineon's Information Security Officer who is backed by the Infineon legal department. In addition to the legal requirements, Infineon's internal data protection regulations ensure that uniform and high data protection standards apply to all companies of the Infineon Group. Employees are obliged to comply with the legal and internal data protection regulations and to protect personal data. All data is secured internally by iArm firewall and AntiVirus. All Microsoft Office and other statistical programs, such as R are updated regularly. Since the confidential data is converted to synthetic data and all orders are anonymised, information flow of sensible company and customer information is counteracted. As the data used is transformed internally, safe transfer is guaranteed by means such as mail encryption and other common company standards. Information security is enforced by regular IT security trainings for employees. Infineon Technologies holds the ISO 27001 certificate by TÜV Nord and is regularly checked in audits. Secure passwords are guaranteed as employees are required to generate high-standard passwords only.

3.3 Data license

The provided datasets feature restricted access on zenodo. That means, in order to receive access, you request access on zenodo. After evaluating your identity and confirming that you are a researcher from the SafeDEED community, access will be granted immediately. However, the dataset is only allowed for non-commerical use only. The credits still belong to Infineon Technologies AG and a copying, displaying or distribution of the dataset to others than the respective researcher that has been granted with access is forbidden and only allowed after approved consent by Infineon Technologies AG. Its use is currently possible for the companies joining the Safe-DEED project. The data is available in this current or might be made available in an alternative form for academic purposes. In any way, it has to be in appliance with the GDPR requirements.

4 Data usage

Only the key synthetic data for Order Lead Times are stored. Data is provided based on business-months. Additional effort for preparing the data is not required, since the file format is Excel. The data may be used in the field of Revenue Management and order behaviour. Raw data will not be provided, only the data obtained after transformation to synthetic data in Excel. Data is retained on a monthly basis when new data is available and stored for up to two years of business-months.

5 Conclusion

In the previous paragraphs, the data composition and creation of the Qualified Synthetic Data for a potential Lead Time Based Pricing approach have been described. Moreover, applied means of data security have been elaborated and usage of the data to scientists has been explained.

The data has been used to find suitable pricing algorithms and will play an important role in finding even better ones in D7.10 as well as testing them together in a simulation method for D7.7. Moreover, QSD has been used in the report on customer segmentation (D7.6), which also required a large amount of data to be shared with partners from other WPs. Since, the original pattern of the data is kept, the results are consistent with the actual results based on the real data. This demonstrates once more the main advantage of QSD.

6 References

[1] Parker, S.P. (1984). McGraw-Hill dictionary of science and engineering. McGraw-Hill, New York.

[2] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L. (2008). Privacy: Theory meets Practice on the Map, Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, IEEE Computer Society, Washington, 277-286.