Grant Agreement Number: 825225

Safe-DEED

www.safe-deed.eu

D7.3 Deliver Qualified Synthetic Data (V1)

Deliverable number	*D7.3*
Dissemination level	*Public*
Delivery date	*Due 29 November 2019*
Status	*Final*
Author(s)	*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
08.11.2019	Alexander Georg	First draft	1.1
21.11.2019	Alexander Georg	Finalised version after review	1.2

Executive summary

Synthetic data make an important contribution to data analysis and research. They are a meaningful representation of real data and anonymized for analysis purpose.

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.

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 the 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 order data are covered in the synthetic data. 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 collected from an algorithm for measuring 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. As Lead Time data is confidential, Qualified Synthetic Data will 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 will initially be approximately 5.5 MB of data, including monthly data for the year-to-date and the previous business year. Data for one Product line is derived and

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

converted for the other Product lines as the distribution is almost identical. Other third-party data is not 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 exemplary. 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 with parameters of up to 50 decimal digits to achieve a high goodness of fit, defined by $R^2 > 0.98$. The synthetic data is then created by inserting values into 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.

	A	B	C	D	E	F	G	H	I	J
1	#	PL	BM	Customer	Order Number	Order Entry Date	Customer Wish Date	Confirmed Delivery Date	Requested Order Lead Time (in d)	Confirmed Order Lead Time (in d)
4454	4453	67	Apr 18	C17	1113652502	03.07.2017	09.04.2018	25.04.2018	280,40	296,41
4455	4454	67	Mai 18	C25	1113195814	25.07.2017	01.05.2018	17.05.2018	280,38	296,39
4456	4455	67	Jan 18	C48	1112559992	21.03.2017	26.12.2017	11.01.2018	280,36	296,37
4457	4456	67	Mai 18	C25	1113195814	25.07.2017	01.05.2018	17.05.2018	280,34	296,35
4458	4457	67	Jun 18	C6	1113437840	12.09.2017	19.06.2018	05.07.2018	280,32	296,34
4459	4458	67	Okt 18	C101	1114041632	09.01.2018	16.10.2018	01.11.2018	280,30	296,32
4460	4459	67	Dez 18	C92	1114435284	20.03.2018	25.12.2018	10.01.2019	280,28	296,30
4461	4460	67	Dez 18	C92	1114435284	20.03.2018	25.12.2018	10.01.2019	280,26	296,29
4462	4461	67	Dez 18	C92	1114435284	20.03.2018	25.12.2018	10.01.2019	280,24	296,27
4463	4462	67	Dez 18	C92	1114435284	20.03.2018	25.12.2018	10.01.2019	280,23	296,25
4464	4463	67	Dez 18	C92	1114435284	20.03.2018	25.12.2018	10.01.2019	280,21	296,23
4465	4464	67	Dez 18	C92	1114435284	20.03.2018	25.12.2018	10.01.2019	280,19	296,22
4466	4465	67	Apr 18	C68	1113110761	07.07.2017	13.04.2018	29.04.2018	280,17	296,20
4467	4466	67	Mrz 18	C18	1112981384	12.06.2017	19.03.2018	04.04.2018	280,15	296,18
4468	4467	67	Mrz 18	C18	1112981384	12.06.2017	19.03.2018	04.04.2018	280,13	296,17
4469	4468	67	Feb 18	C17	1113651277	29.05.2017	05.03.2018	21.03.2018	280,11	296,15
4470	4469	67	Mrz 18	C41	1113041800	23.06.2017	30.03.2018	15.04.2018	280,09	296,13
4471	4470	67	Apr 18	C25	1113195814	25.07.2017	01.05.2018	17.05.2018	280,07	296,11
4472	4471	67	Apr 18	C25	1113195814	25.07.2017	01.05.2018	17.05.2018	280,05	296,10
4473	4472	67	Apr 18	C25	1113195814	25.07.2017	01.05.2018	17.05.2018	280,03	296,08
4474	4473	67	Jun 18	C4	1113531793	29.09.2017	06.07.2018	22.07.2018	280,01	296,06
4475	4474	67	Aug 18	C96	1113805892	21.11.2017	27.08.2018	13.09.2018	279,99	296,05
4476	4475	67	Okt 18	C67	1114041893	09.01.2018	15.10.2018	01.11.2018	279,97	296,03
4477	4476	67	Okt 18	C6	1114116839	23.01.2018	29.10.2018	15.11.2018	279,95	296,01
4478	4477	67	Okt 18	C48	1114146983	27.01.2018	02.11.2018	18.11.2018	279,93	295,99

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

	F	G	H	I	J	K	L	M
1	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)	Order quantity	Product Information (Basic_Type)
4712	19.07.2017	20.04.2018	07.05.2018	275,43	292,01	26,00	1000	P41
4713	19.07.2017	20.04.2018	06.05.2018	275,41	291,99	26,00	1000	P41
4714	30.11.2017	01.09.2018	17.09.2018	275,39	291,97	13,00	1000	P31
4715	26.07.2017	27.04.2018	13.05.2018	275,37	291,96	25,39	22500	P34
4716	26.07.2017	27.04.2018	13.05.2018	275,35	291,94	25,39	47500	P29
4717	07.03.2018	07.12.2018	23.12.2018	275,33	291,92	42,80	7500	P9
4718	22.12.2017	23.09.2018	09.10.2018	275,31	291,90	12,00	2500	P11
4719	07.03.2018	07.12.2018	23.12.2018	275,29	291,89		17500	P9
4720	07.03.2018	07.12.2018	23.12.2018	275,27	291,87	42,80	10000	P9
4721	15.11.2017	17.08.2018	02.09.2018	275,26	291,85	25,39	2500	P29
4722	29.06.2017	31.03.2018	16.04.2018	275,24	291,84	25,39	50000	P27
4723	21.04.2017	21.01.2018	06.02.2018	275,22	291,82	16,00	2000	P204
4724	21.06.2017	23.03.2018	08.04.2018	275,20	291,80	25,39	50000	P34
4725	20.07.2017	21.04.2018	07.05.2018	275,18	291,79	22,89	20000	P151
4726	20.07.2017	21.04.2018	07.05.2018	275,16	291,77	22,89	20000	P151
4727	30.11.2017	01.09.2018	17.09.2018	275,14	291,75	25,39	3990	P76
4728	01.02.2018	03.11.2018	19.11.2018	275,12	291,74	18,21	7500	P36
4729	01.02.2018	03.11.2018	19.11.2018	275,10	291,72	21,02	3000	P53
4730	01.02.2018	03.11.2018	19.11.2018	275,08	291,70	25,39	20000	P43
4731	01.03.2018	01.12.2018	17.12.2018	275,06	291,68	22,89	12000	P146
4732	01.03.2018	01.12.2018	17.12.2018	275,05	291,67	25,39	2500	P32
4733	08.03.2018	08.12.2018	24.12.2018	275,03	291,65	30,33	1134	P133
4734	06.05.2018	05.02.2019	21.02.2019	275,01	291,63	25,39	2500	P76
4735	06.05.2018	04.02.2019	21.02.2019	274,99	291,62	25,39	2500	P76
4736	31.05.2018	01.03.2019	18.03.2019	274,97	291,60	46,00	27000	P144

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 going to be uploaded on Zenodo, an open repository. Findable DOIs will be set, but access will be restricted to researchers and excludes industrial use to prevent spill-overs to direct or indirect competitors. Full data in the final version 2 will be made available in 2020.

Difficulties might occur since the data is very company-sensitive and should not be reproducible. Non-disclosure agreements will be used to avoid the widespread of data.

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.

The data can be found under the URL https://doi.org/10.5281/zenodo.3532890 or the DOI 10.5281/zenodo.3532890.

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 like 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.

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 R. Data is retained on a monthly basis when new data is available and stored for up to two years of business-months. The data is updated throughout the participation of Infineon Technologies AG in the Safe-DEED project. The final version of the data that will be uploaded in May 2020 will contain the synthetic data until April 2020.

5 Conclusion

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

The data will now be used at Infineon as a basis to implement the pricing algorithm that will later be used in Lead Time based Pricing. The data set can also be used by other work packages to test their algorithms. The synthetic data will further be extended to more product lines and business-months and the final version of the data (v2) will be uploaded in May 2020.

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.