Synthetic On Board Diagnostics Data Generation and Evaluation for Vehicle Diagnostic Testing

Veljko Vučinić, Frank Hantschel, Thomas Kotschenreuther
RA Consulting GmbH
Bruchsal, Germany
Email: {v.vucinic, f.hantschel, t.kotschenreuther}@rac.de

***Abstract*—The generation of data plays a vital role in Machine Learning (ML) techniques by providing the foundation for training and improvement of forecast models. As one application area for these models, in-vehicle systems, like vehicle diagnostics, have the potential to enhance the reliability and durability of vehicles by utilizing ML models in the testing phases. However, acquiring a high volume of quality On-Board Diagnostics (OBD) data is time-consuming and poses challenges like the risk of exposing sensitive information. To address this issue, synthetic data generation offers a promising alternative that is already in use in other domains. Thereby, synthetic data allows the exploitation of knowledge found in original data, ensuring the privacy of sensitive data, with less time costs of data acquisition. For this purpose, the research presented in this contribution investigates the use of statistical and ML-based synthetic OBD data generation methods. The models are evaluated with the custom-developed evaluation method that fits the attributes of the OBD data used. Finally, an important result is the successful generation of synthetic OBD data that can be used to enhance the SAE J1699 OBD compliance test, together with tools and insights for models and evaluation.**

***Index Terms*—OBD, SAE J1699, Compliance test, Synthetic Data Generation, Vehicle Testing**

I. Introduction

Vehicle diagnostics has the potential to enhance the reliability and durability of vehicles by utilizing Machine Learning (ML) models in the testing phases. For example, in 2022 85.4 million vehicles have left the productions [1], where every vehicle finished standardized validation tests to check the functionality of crucial vehicle systems. One of those tests is the SAE J1699-3 test used to validate the work and communication of diagnostic systems and main propulsion-related controllers [2]. The SAE J1699-3 test can be extended with the unsupervised ML-based models for identifying early anomalies in vehicle propulsion-related system behavior, based on the knowledge contained in the database used for training such models. As an example, dimension reduction techniques, such as Uniform Manifold Approximation and Projection (UMAP), and self-clustering models, like K-means, can be utilized for the purpose of anomaly detection in vehicle propulsion and diagnostic systems. This has the potential to increase the reliability of vehicles by exploiting anomalistic behavior before a vehicle enters the market.

In this process, access to a large amount of vehicle data is essential for training and enhancing ML methods, where data-based ML models can be utilized for revolutionizing the vehicle diagnostic system [3]. The main challenge is the acquisition of large amounts of high-quality diagnostic data, due to the high resource cost and the risk of exposing sensitive information, e.g. vehicle identification number (VIN) [4].

A solution of these issues can be found in the synthetic data generation. Synthetic data generation models produce artificial datasets and tend to map the features of the original data while safeguarding confidential data and reducing the costs required for data collection [5]. By creating synthetic data from the parts of the SAE J1699 log files, it becomes possible to generate large amounts of OBD data that can be used for training high-quality ML models and testing the system's reliability [6]. With all this in mind, this paper investigates the feasibility of developing SAE J1699 synthetic log data generation models for 54 different OBD parameters, comparing the most relevant statistical and ML-based techniques. Statistical models used in this research are the Kernel Density Estimation (KDE) and Multivariate Gaussian Distribution (MVGD), while ML models included Variational Auto Encoders (VAE) and Generative Adversarial Network (GAN). Moreover, the research presents a customized synthetic generation model evaluation method that fits the specific nature of OBD data.

Synthetic data is already used successfully in a wide range of industries: healthcare [7]–[9], Internet Of Things (IOT) [10], finance [11], manufacturing [12] and many others. For example, recent research [8] utilized the power of synthetic data to eliminate the exposure of sensitive data, thereby ensuring privacy and security for both companies and their customers in the healthcare sector. Furthermore, other research [11] investigated the use of synthetic data generation to fight against financial frauds, among others.

The automotive industry has also embraced the use of synthetic data. Synthetic GPS and temperature data generation for the purpose of implementation and testing of self-learning vehicle functions can be found in [13]. A larger spectrum of generated data is proposed in [14] with a use-case of driver behavior detection, where vehicle speed, engine speed, throttle position, and engine load data were synthetically generated. Both of the mentioned researches focused only on a handful of generated signals. The publications that propose direct use of synthetic data have the purpose of testing for self-learning vehicle comfort function [15] and general system testing [6]. Another potential use case for such data can be the testing of the driver identification function [16], where synthetic data could serve as false driver data.

II. SAE J1699-3 TEST PROCEDURE

The purpose of SAE's recommended practice J1699 is to define test cases for the OBD-II interface on external test equipment (such as an OBD-II scan tool). Therefore, the recommended practice is used to verify vehicle compliance with the applicable standards, such as SAE J1978 [17] and SAE J1979 [18]. The test is done mainly on vehicles during development and production. The J1699-3 represents the test sequence that a vehicle should successfully run in order to guarantee compliance. By definition, the whole test sequence is comprised of separate tests starting from test 5.1 until test 9.23, see Fig. 1. Communication between the OBD-II scan tool and the vehicle is logged and saved in a *.log* file. The communication in the log file is encoded and can be decoded with the use of SAE 1979DA [19]. From the vehicle side, at least one of the relevant onboard electronic control units (ECUs) has to communicate with the scan tool, always by the main engine controller.

Each subtest has a specific purpose, ranging from the basic communication check to the data transfer validation. The operator who implements the test sequence is guided by the standardized software prompts, such as starting or shutting off the engine, disconnecting the sensor, and gradually operating the vehicle until the completion of all contents [20]. The focus of this research will be on the tests 5.6 and 5.10. Both of those tests have the purpose of verifying the service $01 data, where 5.6 is done with the engine off and 5.10 with the engine running. Service $01 represents the current propulsion- and emission-related data (i.e. throttle position, engine speed, oxygen sensor, engine oil temperature). In total, there are over 200 diagnostic parameters that can be available for a vehicle [19], but on average less than 100 are available on a vehicle. These two tests provide a base for potential ML enhancement of SAE J1699-3 due to the generated snapshots of vehicle diagnostic data. The snapshots can provide information about anomalistic behavior of a vehicle, that can be exploited with the proper use of clustering and dimension reduction algorithms.

III. METHODOLOGY

In order to successfully generate a large amount of synthetic SAE J1699 OBD data, this paper followed the Knowledge Discovery in Databases (KDD) process [21]. The methodology contains several precise steps ranging from data acquisition to data processing and finally synthetic data generation, see Fig. 2.

A. Data acquisition

Data acquisition is a critical first step in the synthetic data generation process, as it provides the foundation for creating quality, representative, and useful synthetic datasets. For this purpose, the relevant parts of the SAE J1699-3 test were done on 57 different vehicles and therefore collected the same amount of log files including the snapshots from the engine off and engine running OBD data. The data acquisition is done using the Silver Scan-Tool™ software tool, with the support of a VSI-2534 as the physical interface (passthrough device).

B. Data preprocessing

The data preprocessing took place after the acquisition. First, the data was extracted from the encoded raw log files, and the data was converted from the hexadecimal to the decimal format. After the extraction of the data, classical preprocessing techniques followed, including removing outliers and filtering. Following, min-max normalization is applied to ensure that all features contribute uniformly to the accuracy and robustness of the generation model. Finally, the data is split into train and test datasets with a ratio of 85%:15%. In order to keep the uniformity of the data generation models, OBD parameters that are engine architecture specific (i.e. found only in diesel or petrol engines) were removed. With this in mind, the models trained are not exclusive to any engine or vehicle class.

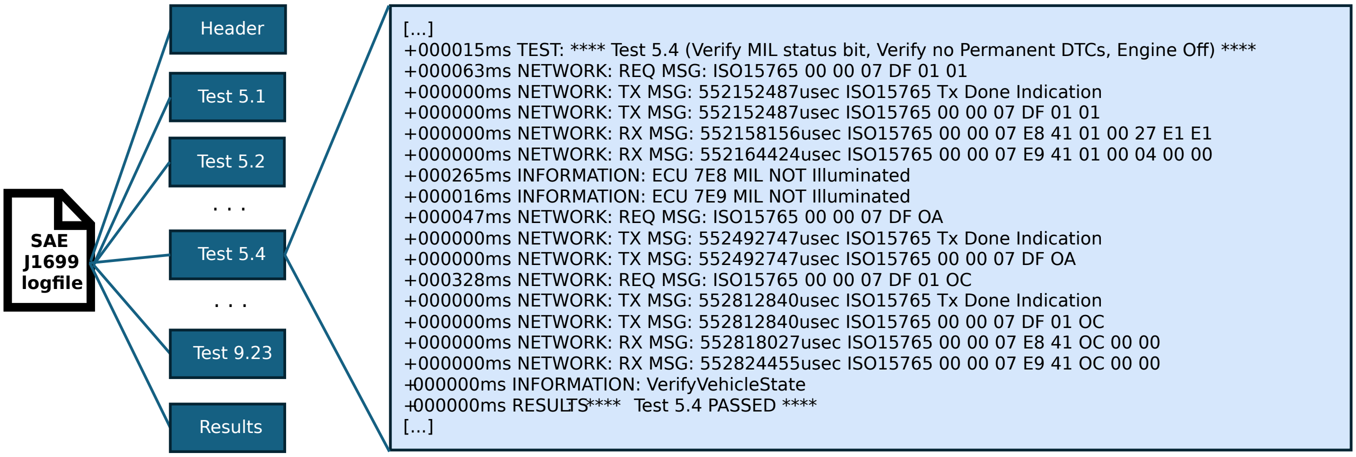
C. Generative models

This work evaluates the four most popular generative models for OBD synthetic data generation. The first two are statistic-based: Kernel Density Estimation (KDE) and Multivariate Gaussian Distribution (MVGD). Additionally, two ML-based models are considered, namely Generative Adversive Networks (GAN) and Variational Auto Encoders (VAE).

KDE represents a non-parametric method used to estimate the probability density function of a random variable based on a finite data sample. This method can be applied to the original dataset of preprocessed SAE J1699 data to estimate the underlying probability density function. During this process, it is necessary to choose a kernel function and bandwidth to smooth the data. The data generation with KDE is done by sampling data points from the estimated distribution of a density function [22].

MVGD is an extended single-variable Gaussian distribution to higher dimensions. It has the capacity to simulate the joint probability distribution of several variables, and therefore, it is a suitable candidate to produce synthetic data. The procedure entails sampling new data points from the multivariate Gaussian distribution specified by the estimated mean vector and covariance matrix from the actual data. This guarantees that the synthetic data generated preserves the original dataset's statistical properties, including means, variances, and correlations [23].

Both GANs and VAEs are an approach to generative modeling using deep learning methods. On one hand, GANs consist of two neural networks that compete against each other. One is a generator that creates fake data samples, and the other is a discriminator that tries to distinguish the original from the fake data. The final goal is to train the generator to produce data that the discriminator model cannot distinguish from the original [24]. Lastly, VAEs are a generative model that combines principles from autoencoders and variational inference. They are designed to learn a probabilistic mapping from input data to a latent space and then generate new data



```
[...]
+000015ms TEST: **** Test 5.4 (Verify MIL status bit, Verify no Permanent DTCs, Engine Off) ****
+000063ms NETWORK: REQ MSG: ISO15765 00 00 07 DF 01 01
+000000ms NETWORK: TX MSG: 552152487usec ISO15765 Tx Done Indication
+000000ms NETWORK: TX MSG: 552152487usec ISO15765 00 00 07 DF 01 01
+000000ms NETWORK: RX MSG: 552158156usec ISO15765 00 00 07 E8 41 01 00 27 E1 E1
+000000ms NETWORK: RX MSG: 552164424usec ISO15765 00 00 07 E9 41 01 00 04 00 00
+000265ms INFORMATION: ECU 7E8 MIL NOT Illuminated
+000016ms INFORMATION: ECU 7E9 MIL NOT Illuminated
+000047ms NETWORK: REQ MSG: ISO15765 00 00 07 DF OA
+000000ms NETWORK: TX MSG: 552492747usec ISO15765 Tx Done Indication
+000000ms NETWORK: TX MSG: 552492747usec ISO15765 00 00 07 DF OA
+000328ms NETWORK: REQ MSG: ISO15765 00 00 07 DF 01 OC
+000000ms NETWORK: TX MSG: 552812840usec ISO15765 Tx Done Indication
+000000ms NETWORK: TX MSG: 552812840usec ISO15765 00 00 07 DF 01 OC
+000000ms NETWORK: RX MSG: 552818027usec ISO15765 00 00 07 E8 41 OC 00 00
+000000ms NETWORK: RX MSG: 552824455usec ISO15765 00 00 07 E9 41 OC 00 00
+000000ms INFORMATION: VerifyVehicleState
+000000ms RESULTS****  Test 5.4 PASSED ****
[...]
```

Fig. 1. Representation of SAE J1699-3 test sequence with a part of a log file.

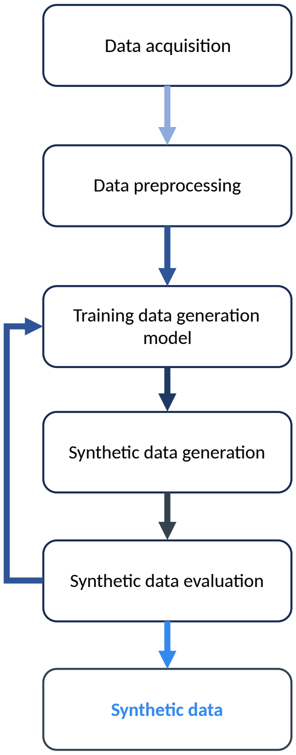

Data acquisition
Data preprocessing
Training data generation model
Synthetic data generation
Synthetic data evaluation
Synthetic data

Fig. 2. Synthetic OBD data generation methodology

samples from this latent space. Encoders in VAEs are used to map input data to a lower-dimensional, latent space, and decoders are used to map data points back to the original space. After training both the encoder and decoder, the VAE model can generate artificial data by sampling random points through trained blocks [25].

D. Evaluation

Data generated using generative models needs to be transparently evaluated and compared. For this purpose, the research presented in this paper proposes a unique evaluation method that fits the specific nature of OBD data and the SAE J1699 case. The specific evaluation methodology is built upon two well-accepted statistical tests and serves the purpose of evaluating how close are two datasets by their statistical features. The first one is the Kolmogorov-Smirnov (KS) test, which evaluates the cumulative distribution function between synthetic and original data. It is used to compare distributions of two sets of data [26]. The outputs of KS tests are the test statistic variable D and the p-value p_{KS}. The second is the T-test that takes into consideration the mean difference between the datasets [27]. The outputs of this test are specific t-statistic t and T-test p-value p_T. Based on these four variables from both tests, we created an evaluation for grading of generated synthetic OBD data, see Equation 1.

$$G = \frac{1}{n_{PID}} \sum_{k=1}^{n_{PID}} \left[w_{KS} \Big(1 - D(k) \Big) + \frac{1 + p_{KS}(k) + p_T(k)}{w_T |t(k)|} \right] \tag{1}$$

Total synthetic dataset grade G is represented as an average grade of all of its OBD parameters (PIDs). Lower values of parameters D and t indicate similar distributions and mean values between synthetic and real data, elevating the grade in Equation 1. On the other hand, small values of both test p-values suggest lower statistical significance of observed correlation of distribution and mean for different tests, and therefore punish the grade of the synthetic dataset. In the case of this work total number of OBD parameters is $n_{PID} = 54$. Each OBD parameter has its resemblance to the parameter from the original data, represented through statistics tests parameters $D(k)$, $p_{KS}(k)$, $t(k)$, and $p_T(k)$. The goal of the specific evaluation method is to be transparent and easily understandable. With this in mind, the purpose of the weights w_{KS}, w_T is to keep the total grade in the range of 1 to 10, where 1 means that synthetic data poorly represents the original dataset, and 10 indicates high statistical similarity towards original data. For this research, exact weight values are $w_{KS} = 8$ and $w_{KS} = 7$.

IV. Synthetic OBD data generation

Synthetic data is artificially generated data which mimics the real-world data. The synthetic data generation in this work took place separately for the engine off (test 5.6) and engine running (test 5.10) parts of the SAE J1699. The complete list of the OBD parameters considered for data generation in both cases is listed together with their standardized PIDs in Table I.

TABLE I
List of OBD parameters considered for synthetic data generation.

PID	Name	Description
01	MIL	Malfunction Indicator Lamp Status
04	LOAD_PCT	Calculated LOAD Value
05	ECT	Engine Coolant Temperature
06	SHRTFT1	Short Term Fuel Trim - Bank 1
07	LONGFT1	Long Term Fuel Trim – Bank 1
0B	MAP	Intake Manifold Absolute Pressure
0C	RPM	Engine RPM
OD	VSS	Vehicle Speed Sensor
0F	IAT	Intake Air Temperature
10	MAF	Air Flow Rate
11	TP	Absolute Throttle Position
15	O2Sv12	Oxygen Sensor Output Voltage
15	SHRTFT12	Oxygen Sensor 2 Short term fuel trim
1C	OBDSUP	OBD requirements of vehicle
1F	RUNTM	Time Since Engine Start
21	MIL_DIST	Distance Traveled While MIL is Activated
23	FRP	Fuel Rail Pressure
24	O2SV11	Oxygen Sensor Voltage - Bank 1, Sensor 1
2E	EVAP_PCT	Commanded Evaporative Purge
2F	FLI	Fuel Level Input
30	WARM_UPS	Number of warm-ups since DTCs cleared
31	CLR_DIST	Distance traveled since DTCs cleared
33	BARO	Barometric Pressure
34	LAMBDA11	Equivalence Ration - Bank 1, Sensor 1
34	O2Sc11	Oxygen Sensor Current - Bank 1, Sensor 1
3C	CATEMP11	Catalyst temperature Bank 1 Sensor 1
42	VPWR	Control module voltage
43	LOAD_ABS	Absolute Load Value
44	LAMBDA	Fuel/Air Commanded Equivalence Ratio
45	TP_R	Relative Throttle Position
46	AAT	Ambient air temperature
47	TP_B	Absolute Throttle Position B
49	APP_D	Accelerator Pedal Position D
4A	APP_E	Accelerator Pedal Position E
4C	TAC_PCT	Commanded Throttle Actuator Control
53	EVAP_VPA	Absolute Evap System Vapor Pressure
56	LGSO2FT1	Long Term Secondary O2 Sensor Fuel Trim
5C	EOT	Engine Oil Temperature
5E	FUEL_RATE	Engine Fuel Rate
62	TQ_ACT	Actual Engine - Percent Torque
63	TQ_REF	Engine Reference Torque
67	ECT_1	Engine Coolant Temperature 1
67	ECT_2	Engine Coolant Temperature 2
68	IAT_11	Intake Air Temperature - Bank 1, Sensor 1
68	IAT_12	Intake Air Temperature - Bank 1, Sensor 2
73	EP_1	Exhaust Pressure Sensor Bank 1
78	EGT11	Exhaust Gas Temperature - Bank 1, Sensor 1
78	EGT12	Exhaust Gas Temperature - Bank 1, Sensor 2
8E	TQ_FR	Engine Friction - Percent Torque
9D	ENG_FUEL_RATE	Engine Fuel Rate
9D	VEH_FUEL_RATE	Vehicle Fuel Rate
9E	EXH_RATE	Engine Exhaust Flow Rate
A4	GEAR_ACT	Actual Transmission Gear
A6	ODO	Odometer

The synthetic data generation is done using the four methods disclosed in the section III-C for three specific cases, 100, 500, and 1000 newly generated SAE J1699 tests 5.6 and 5.10. With this in mind, the initial SAE J1699 log file OBD dataset for future ML models of 57 vehicles has been extended with 100, 500, and 1000, novel synthetic vehicles, respectively for each scenario. The knowledge from the original data is replicated for the quantities of synthetic data that can make ML models perform better after training phases, in comparison to the size of the original data. Evaluation of three scenarios for all four generation models is done to grade the quality of synthetic data with regard to the data quantity. All four generative models were implemented for various parameters, and finally, the best combination of parameters was selected for evaluation. For the KDE model, the best output data had a bandwidth of 0.1 and a uniform (top hat) kernel function. The GAN model which provided the best result was implemented with Binary Cross-Entropy (BCE) as loss function, Adam as an optimizer, latent dimension of 100, learning rate $1e{-4}$, number of epochs for training 5000 and batch size of 32. The VAE was selected with BCE loss function, Adam as optimized, latent dimension of 10, batch size of 32, learning rate $1e{-4}$ and 5000 epochs. The MVGD model has no tunable parameter for synthetic data generation.

The visualized results of the synthetically generated data from the best models for engine off and engine on scenarios are shown in the sections A and B, respectively. The generated screenshots of OBD data (blue data points) are combined with the original data from the J1699 log files (black data points). The horizontal axis represents the index of one J1699 5.6 and 5.10 test.

V. Synthetic data evaluation

Evaluation of synthetic data is a necessary process to check the quality and usability of the data generated. Data quality refers to the accuracy, validity, uniqueness, and consistency of synthetic data. A specialized evaluation method, presented in the section III-D, is applied to all generated data. Evaluation is done separately to the generated engine off data (test 5.6) and engine running data (test 5.10).

A. Evaluation of engine off data

Engine off data has certain parameters that are constant for all vehicles, due to the nature of the test 5.6 (i.e. Engine RPM, Calculated LOAD Value, etc.). Therefore, it is expected that the models are able to recognize the structure of the engine off data better than in another case. The results show that the best model in this case is KDE, slightly in front of MVGD, see Fig. 3. Generated engine off data from the KDE model has indeed the closest structure to the original data for almost all OBD parameters, as visualized in Figure 7, in comparison to other models, Figures 5, 6, and 8. For the case of generated 100 tests, VAE shows high quality with a grade of 6.9, slightly behind the two statistical methods. As the generated data size increases, VAE drastically drops grade, while with other methods GAN keeps close to a constant level of lower synthetic data quality. In the case of statistic methods, a slight increase in grade exists as the data size increases. The small size of the original data (57 tests) can be the cause of the domination of statistical among ML-based models.

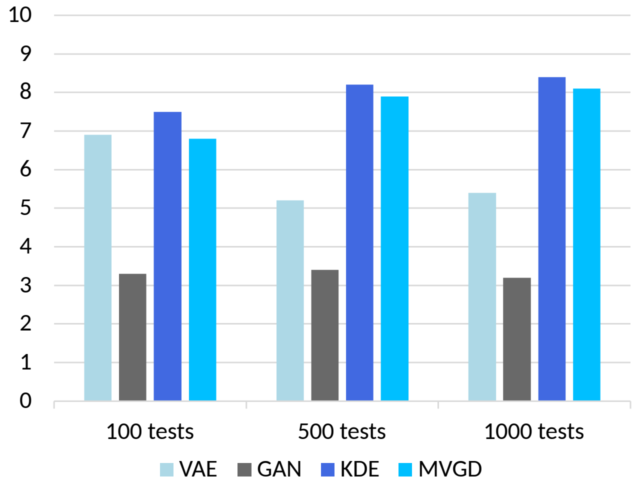


Fig. 3. Evaluation results for the case of engine off OBD data (Test 5.6).

B. Evaluation of engine running data

In the case of test 5.10, the structure of the data is more complex than in the previous case, due to the running state of the engine. This means that the screenshot of OBD parameters done in test 5.10 is more prone to randomness and dependent on the time instance when the screenshot is done. Nevertheless, the results for the synthetic data generation of an engine on data are quite similar to the first case, see Fig. 4. Still, domination is found in statistical models, while MVGD is slightly behind the KDE. In this case, VAE created much less similar data to the original, in comparison to the case of engine off. On the other hand, the GAN model performs better but is still out of the range of statistical models. The high quality of the generated engine running data using KDE is evident in Figure 11. The results of engine running generated data of other models are presented in Figures 9, 10, and 12

VI. Conclution and Future Work

The presented research investigated the possibility of synthetic data generation for the case of OBD data from the SAE J1699 log files. The purpose is to generate data that would in quantity be sufficient for training ML models for various vehicle diagnostic applications and would be a match in quality compared to the real-world data. By that, additional costs of acquiring large quantities of real-world data and the risk of exposing sensitive information can be avoided.

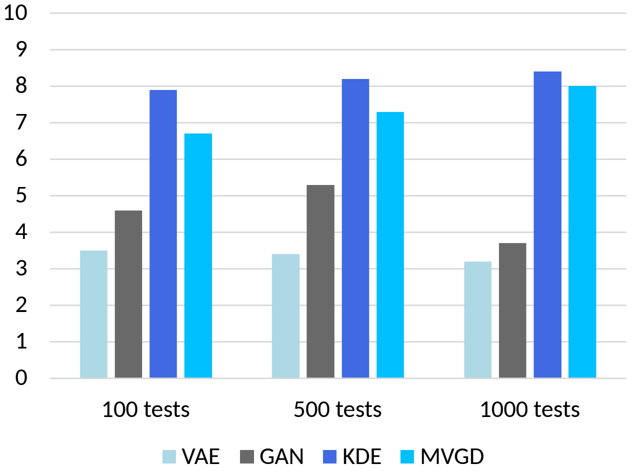


Fig. 4. Evaluation results for the case of engine running OBD data (Test 5.10).

The results showed that it is possible to create quality synthetic OBD data from the small initial dataset of 57 different tests in large quantities. The generated data sizes in this research were as high as 1000 new tests. The best result outputted the KDE model in both use cases of engine off (test 5.6) and engine running (test 5.10) vehicle states. The synthetic data was evaluated using a specifically designed evaluation method that fits the features of OBD data. Results show the dominance of statistical generation models (KDE and MVGD) over ML-based (VAE, GAN). The reason can be found in the small training dataset, which is crucial for the performance of ML-based generation models.

Future work is aimed to investigate the usability of such generated data with the goal of AI-based applications that can append standardized SAE J1699-3 compliance tests, therefore increasing the reliability of future vehicles.

Acknowledgment

The authors would like to thank and recognize the effort from Manish Varri for the support of this research work.

References

[1] ACEA, *World motor vehicle production*, https://www.acea.auto/figure/world-motor-vehicle-production/, Accessed: 2010-09-30.

[2] SAE International, *J1699-3: Vehicle OBD II compliance test cases*, https://www.sae.org/standards/content/j1699/3_202104/, 2021.

[3] M. Al-Zeyadi, J. Andreu-Perez, H. Hagras, *et al.*, "Deep learning towards intelligent vehicle fault diagnosis," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–7.

[4] M. Cheah, S. Haynes, and P. Wooderson, "Smart vehicles: The data privacy smog," in *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, IEEE, 2018, pp. 82–87.

[5] T. E. Raghunathan, "Synthetic data," *Annual review of statistics and its application*, vol. 8, no. 1, pp. 129–140, 2021.

[6] G. Soltana, M. Sabetzadeh, and L. C. Briand, “Synthetic data generation for statistical testing,” in *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, IEEE, 2017, pp. 872–882.

[7] J. Dahmen and D. Cook, “Synsys: A synthetic data generation system for healthcare applications,” *Sensors*, vol. 19, no. 5, p. 1181, 2019.

[8] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson, and F. Mahmood, “Synthetic data in machine learning for medicine and healthcare,” *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 493–497, 2021.

[9] J. Hyun, Y. Lee, H. M. Son, *et al.*, “Synthetic data generation system for ai-based diabetic foot diagnosis,” *SN Computer Science*, vol. 2, no. 5, p. 345, 2021.

[10] J. W. Anderson, K. E. Kennedy, L. B. Ngo, A. Luckow, and A. W. Apon, “Synthetic data generation for the internet of things,” in *2014 IEEE International Conference on Big Data (Big Data)*, IEEE, 2014, pp. 171–176.

[11] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso, “Generating synthetic data in finance: Opportunities, challenges and pitfalls,” in *Proceedings of the First ACM International Conference on AI in Finance*, 2020, pp. 1–8.

[12] D. Libes, D. Lechevalier, and S. Jain, “Issues in synthetic data generation for advanced manufacturing,” in *2017 IEEE International Conference on Big Data (Big Data)*, IEEE, 2017, pp. 1746–1754.

[13] M. Stang, M. G. Marquez, and E. Sax, “Cagen-context-action generation for testing self-learning functions,” in *Human Interaction, Emerging Technologies and Future Applications IV: Proceedings of the 4th International Conference on Human Interaction and Emerging Technologies: Future Applications (IHIET–AI 2021), April 28-30, 2021, Strasbourg, France 4*, Springer, 2021, pp. 12–19.

[14] E. Ucuzova, E. Kurtulmaz, F. G. Yavuz, H. Karacan, and N. E. Şahın, “Synthetic canbus data generation for driver behavior modeling,” in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, IEEE, 2021, pp. 1–4.

[15] M. Stang, L. Seidel, V. Vučinić, and E. Sax, “Exploring metamorphic testing for self-learning functions with user interactions,” *Human Interaction and Emerging Technologies (IHIET-AI 2024): Artificial Intelligence and Future Applications*, vol. 120, no. 120, 2024.

[16] V. Vučinić, L. Seidel, M. Stang, and E. Sax, “USID-unsupervised identification of the driver for vehicle comfort functions,” *Human Interaction and Emerging Technologies (IHIET-AI 2024): Artificial Intelligence and Future Applications*, vol. 120, no. 120, 2024.

[17] SAE International, *J1978: OBD-II scan tool*, https://www.sae.org/standards/content/j1978_202205/, 2022.

[18] SAE International, *J1979: E/E diagnostic test modes*, https://www.sae.org/standards/content/j1979_201202/, 2012.

[19] SAE International, *J1979-DA: Digital annex of E/E diagnostic test modes*, https://www.sae.org/standards/content/j1979da_201406/, 2014.

[20] L. Wang, X. Zou, H. Qin, and P. Geng, “Design of OBD function test on production vehicle (pve),” in *E3S Web of Conferences*, EDP Sciences, vol. 268, 2021, p. 01 047.

[21] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, “Knowledge discovery in databases: An overview,” *AI magazine*, vol. 13, no. 3, pp. 57–57, 1992.

[22] Y.-C. Chen, “A tutorial on kernel density estimation and recent advances,” *Biostatistics & Epidemiology*, vol. 1, no. 1, pp. 161–187, 2017.

[23] C. B. Do, “The multivariate gaussian distribution,” *Section Notes, Lecture on Machine Learning, CS*, vol. 229, 2008.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[25] S. I. Nikolenko, *Synthetic data for deep learning*. Springer, 2021, vol. 174.

[26] G. Fasano and A. Franceschini, “A multidimensional version of the kolmogorov–smirnov test,” *Monthly Notices of the Royal Astronomical Society*, vol. 225, no. 1, pp. 155–170, 1987.

[27] T. K. Kim, “T test as a parametric statistic,” *Korean journal of anesthesiology*, vol. 68, no. 6, pp. 540–546, 2015.

[28] M. K. Varri, “Development and comparison of synthetic data generation models for on-board diagnostic data,” M.S. thesis, Karlsruhe University of Applied Sciences, 2024.

APPENDIX

A. Original and synthetic generated for Test 5.6

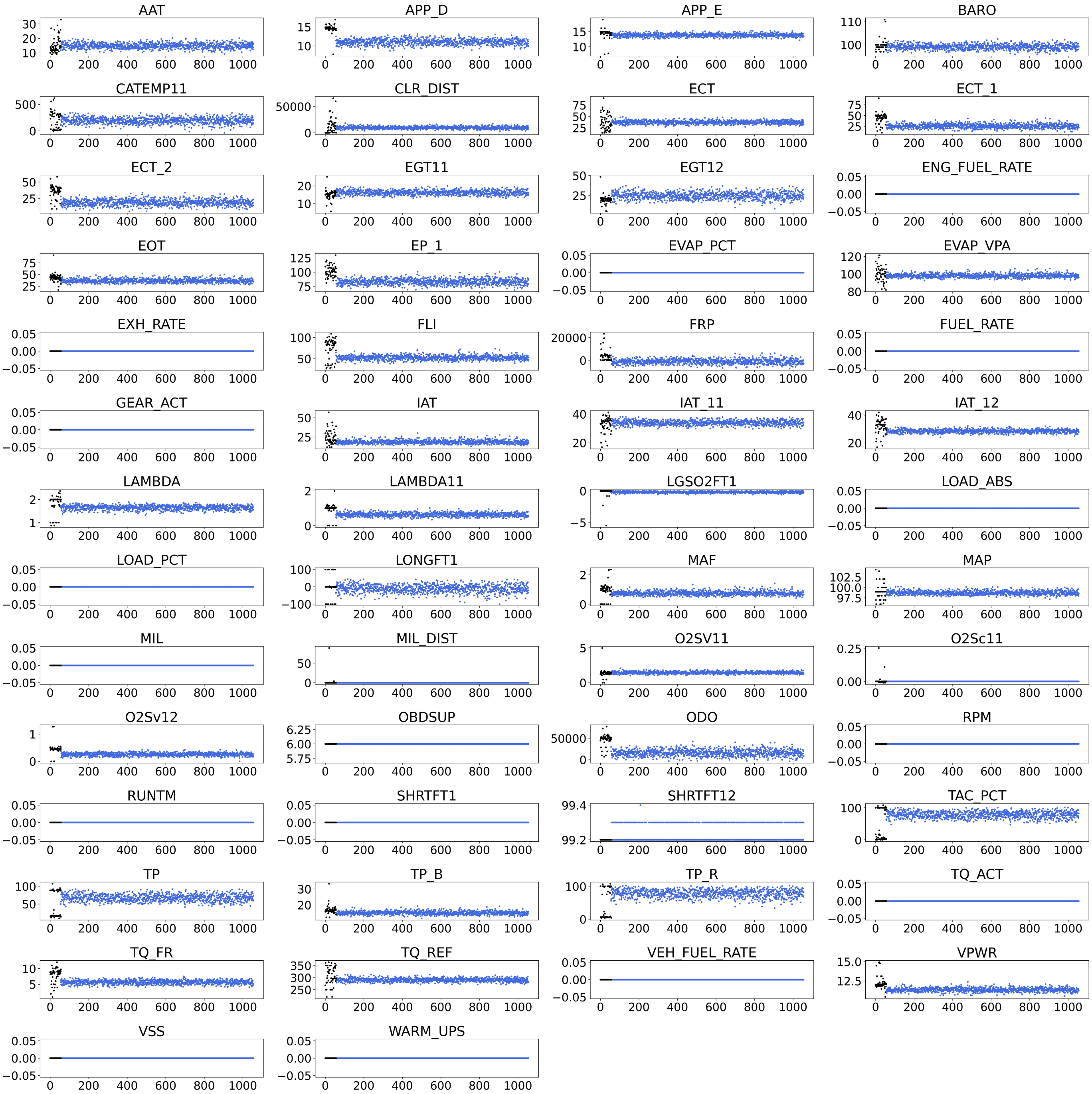


Fig. 5. Comparison of original data (black) and GAN generated data (blue) for the case of 1000 tests and engine off (test 5.6) with all 54 OBD parameters. Each point represents a parameter value from one vehicle.

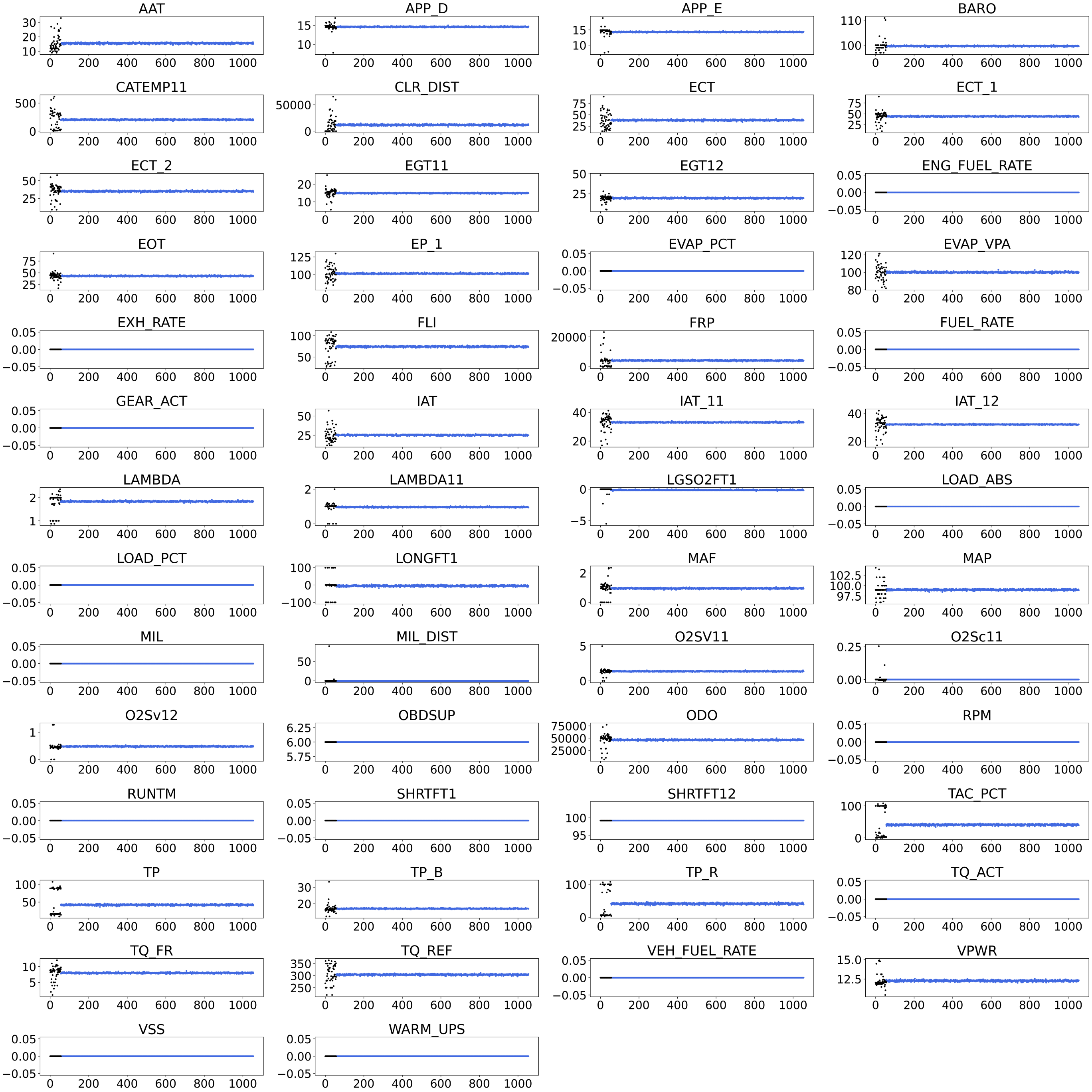


Fig. 6. Comparison of original data (black) and VAE generated data (blue) for the case of 1000 tests and engine off (test 5.6) with all 54 OBD parameters. Each point represents a parameter value from one vehicle.

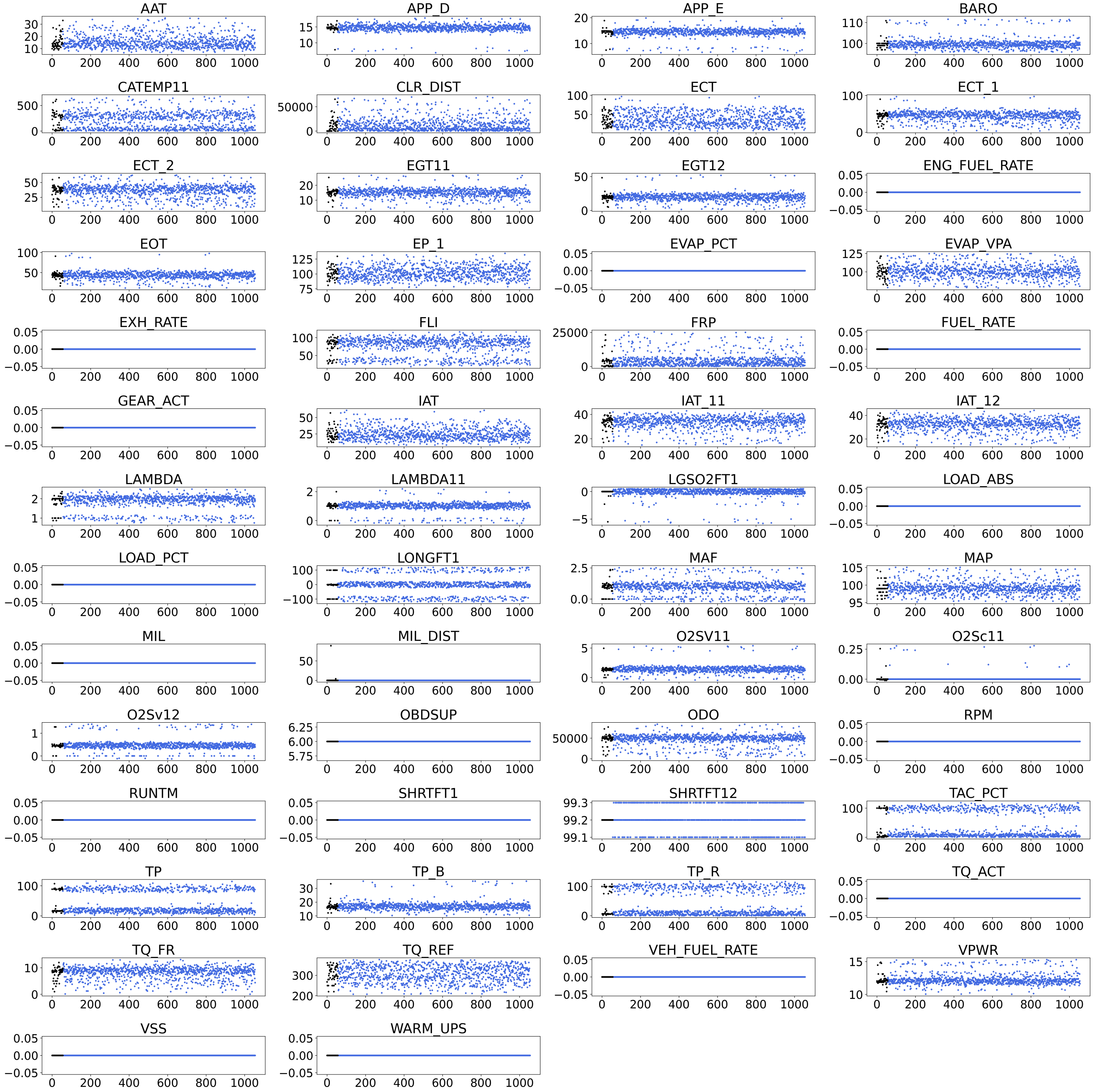


Fig. 7. Comparison of original data (black) and KDE generated data (blue) for the case of 1000 tests and engine off (test 5.6) with all 54 OBD parameters. Each point represents a parameter value from one vehicle.

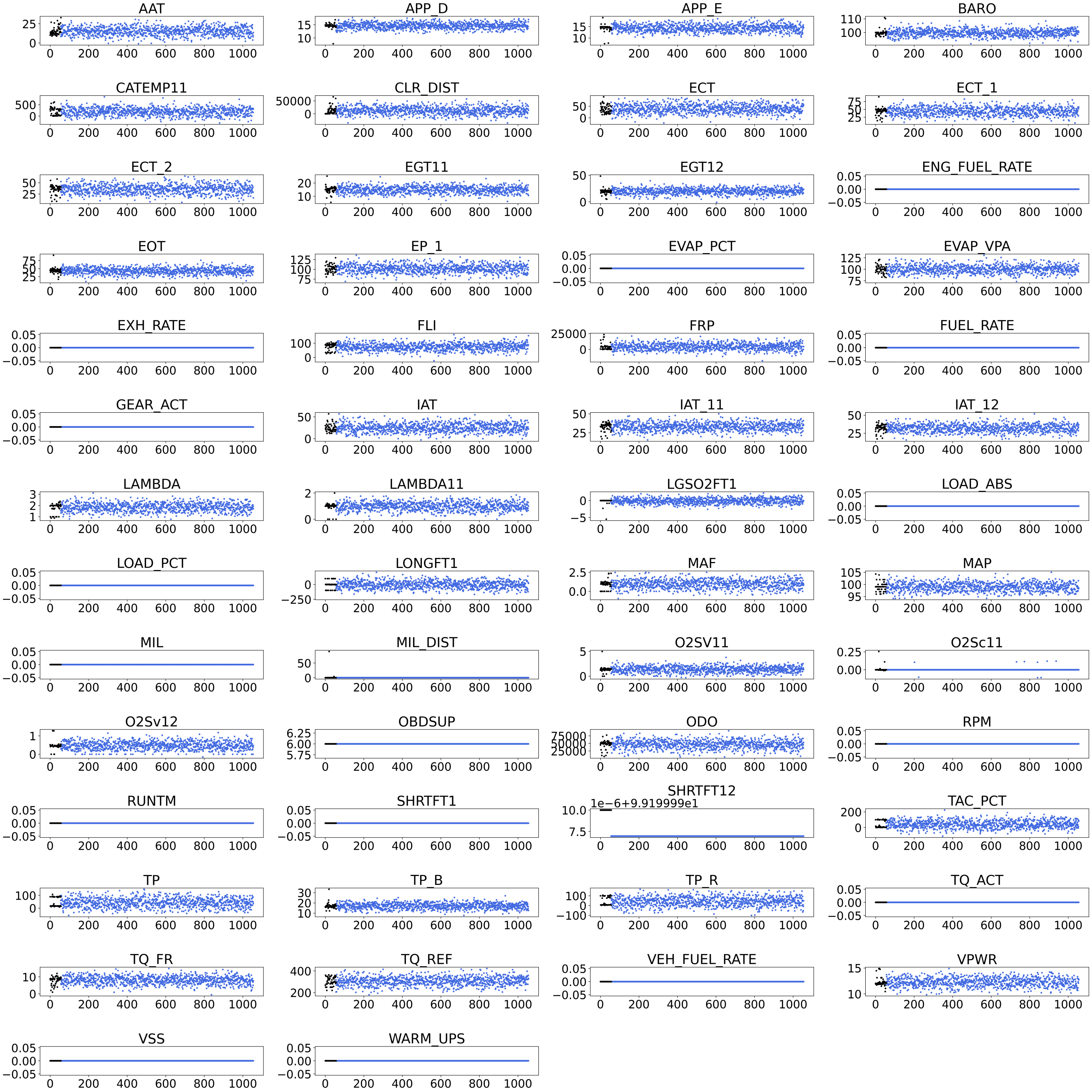


Fig. 8. Comparison of original data (black) and MVGD generated data (blue) for the case of 1000 tests and engine off (test 5.6) with all 54 OBD parameters. Each point represents a parameter value from one vehicle.

B. *Original and synthetic generated for test 5.10*

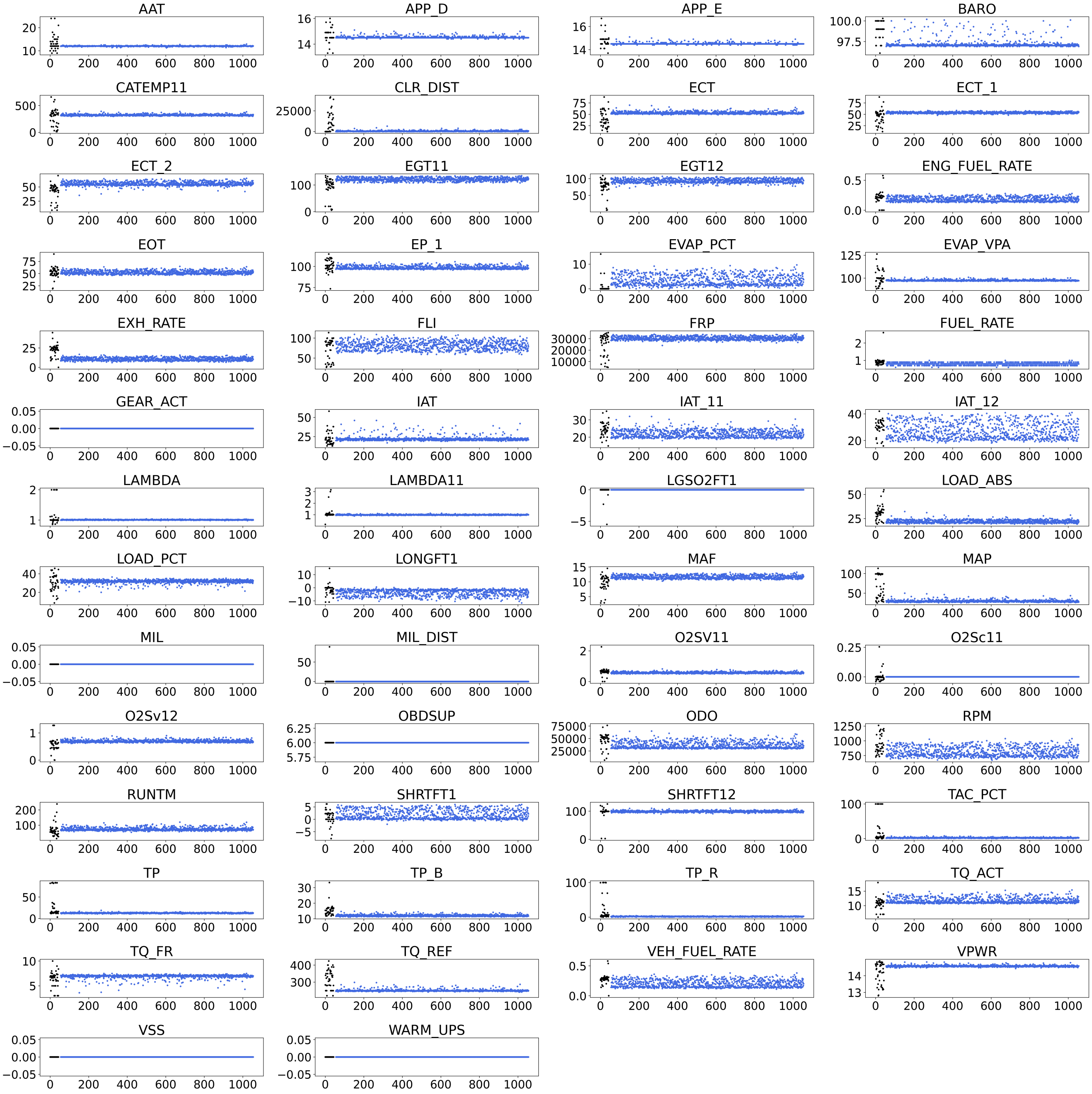


Fig. 9. Comparison of original data (black) and GAN generated data (blue) for the case of 1000 tests and engine on (test 5.10) with all 54 OBD parameters. Each point represents a parameter value from one vehicle.

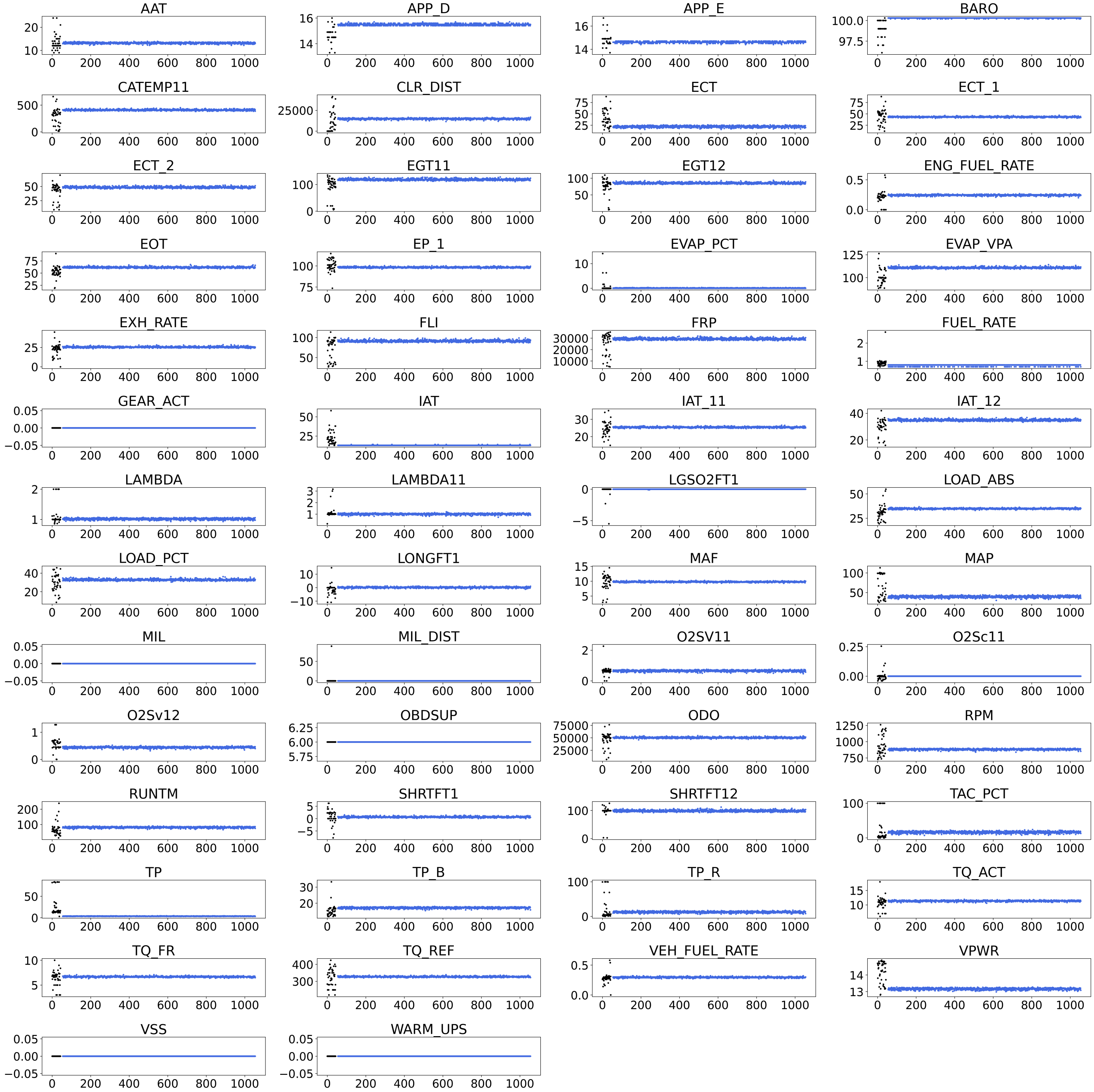


Fig. 10. Comparison of original data (black) and VAE generated data (blue) for the case of 1000 tests and engine on (test 5.10) with all 54 OBD parameters. Each point represents a parameter value from one vehicle.

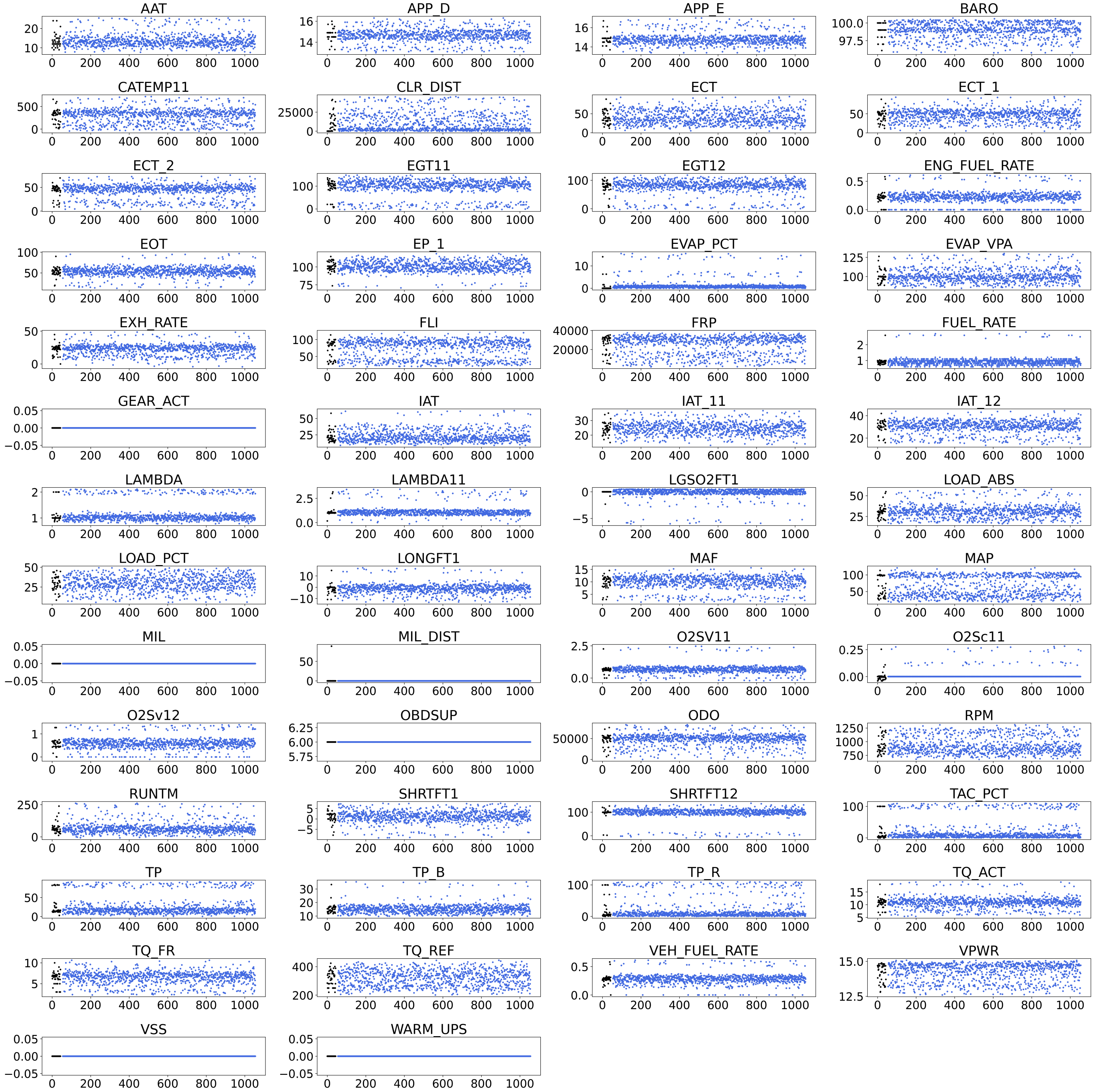


Fig. 11. Comparison of original data (black) and KDE generated data (blue) for the case of 1000 tests and engine on (test 5.10) with all 54 OBD parameters. Each point represents a parameter value from one vehicle.

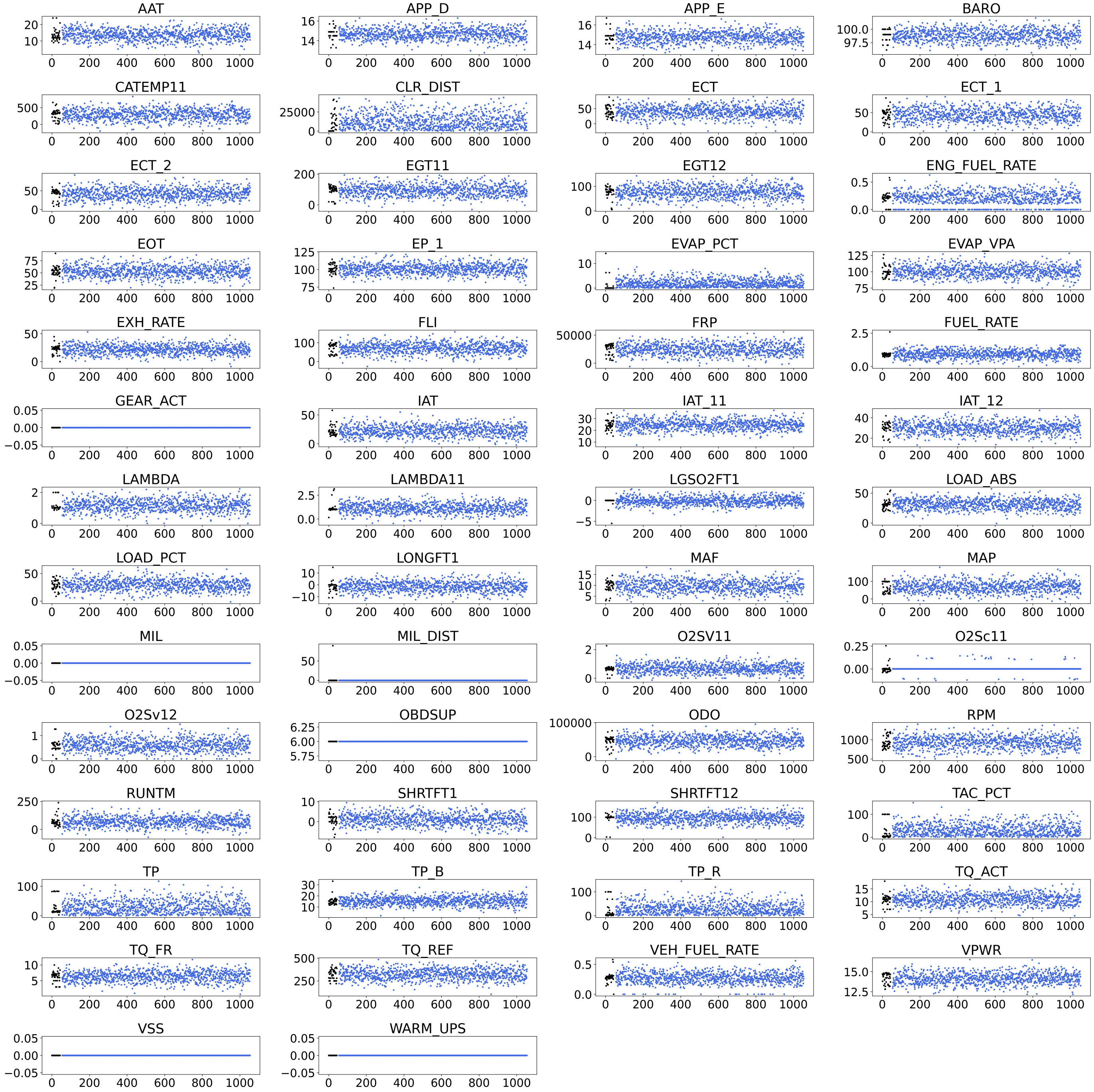


Fig. 12. Comparison of original data (black) and MVGD generated data (blue) for the case of 1000 tests and engine on (test 5.10) with all 54 OBD parameters. Each point represents a parameter value from one vehicle.