https://doi.org/10.14669/AM/189172

GAUSSIAN PROCESS REGRESSION AS A PRECRASH VELOCITY DETERMINATION METHOD– SUBCOMPACT VEHICLE CLASS

FILIP TUROBOŚ[1], ADAM MROWICKI[2], PIOTR KONIECZNY[3], SZYMON MADZIARA[4], ŁUKASZ STAJUDA[5], BRANISLAV ŠARKAN[6], PRZEMYSŁAW KUBIAK[7], DYMYTRO LEVCHENKO[8], NICOLE MEISNER[9]

Abstract

The following paper presents an innovative approach to determining vehicle precrash velocity when hitting an immovable obstacle facing forward. Precrash velocity is necessary in order to perform a crash reconstruction. It is needed for the time–space analysis of the events, as well as to assess crash mitigation and to evaluate drivers' technique and tactics. For this task, the authors are using Gaussian Process Regression (GPR). Such an approach offers a number of advantages over the currently used methods that prove to be outdated when considering modern vehicles. The mathematical model was trained on

[1] Institute of Mathematics, Lodz University of Technology, Poland, e-mail: filip.turobos@p.lodz.pl, ORCID: 0000-0002-5782-6159

[2] Institute of Vehicles and Construction Machinery Engineering, Warsaw University of Technology, Poland, e-mail: adam.mrowicki.dokt@pw.edu.pl, ORCID: 0000-0002-7865-099X

[3] Institute of Mathematics, Lodz University of Technology, Poland, e-mail: 242616@edu.p.lodz.pl, ORCID: 0000-0003-4755-744X

[4] Institute of Vehicles and Construction Machinery Engineering, Warsaw University of Technology, Poland, e-mail: szymon.madziara.dokt@pw.edu.pl, ORCID: 0000-0002-6193-9260

[5] Division of Ecotechnics, Lodz University of Technology, Poland, e-mail: lukasz.stajuda@lodzkie.pl, ORCID: 0009-0008-5086-9245

[6] Faculty of Operation and Economics of Transport and Communications, University of Žilina, Slovak Republic, e-mail: branislav.sarkan@uniza.sk, ORCID: 0000-0002-5036-9223

[7] Division of Ecotechnics, Lodz University of Technology, Poland, e-mail: przemyslaw.kubiak@p.lodz.pl, ORCID: 0000-0003-0225-9609

[8] Division of Ecotechnics, Lodz University of Technology, Poland, e-mail: dymytro.levchenko@p.lodz.pl, ORCID: 0000-0003-0766-4371

[9] Department of Mathematics, University of Warsaw, Poland, e-mail: meisnernicole@gmail.com, ORCID: 0009-0006-3798-3196

https://doi.org/10.14669/AM/189172

a database shared by the National Highway Traffic Safety Administration. This database covers a large number of crash tests of different kind, however authors focus on frontal collisions of the subcompact car class. Due to low accuracy of linear methods used up till now, Authors developed an innovative approach to determine the EES parameter utilizing Gaussian process regression. The newly developed method is an effective and accurate way to determine the vehicle's velocity and shows promising results, as is demonstrated in this paper.

Keywords: car crash reconstruction; EES; Gaussian Process Regression

1. Introduction

Vehicle crash analysis is an important branch of forensic science. The outcome of such investigation has proven to be vital evidence in court trials; therefore, the accuracy of the analysis is of upmost importance. Currently, there are three major approaches to crash velocity determination: graphical, comparative and analytical. Their reliability and accuracy can be questionable, especially when considering modern cars that utilise advanced materials or special structures. Additionally, the introduction of electric vehicles [2, 6] to the market forces the development of new methods suitable for different structural characteristics.

The linear models used thus far do not provide the proper accuracy when determining precrash velocity $\mathrm{V_t}$, which is a critical factor during car-crash reconstruction [11]. The new method takes advantage of the nonlinear relationship between velocity V_t, mass m and deformation coefficient C_s which is the arithmetic mean of deformation depth. The currently used methods assume the relationship between equivalent energy speed (EES) [1, 8] and the $\mathrm{C_s}$ coefficient to be linear. This was done to simplify the calculations and decrease the required computational power as these methods were created in the nineteen-eighties.

The NHTSA database covers crash test results of, among other types, frontal collisions. The linear approach revolves around the energy of the crash and assumes inelasticity of collisions [5]. Based on the literature, the threshold of elasticity is set at 11 km/h. Such simplification further decreases the reliability of energy methods.

The authors decided to develop a new approach to precrash velocity determination using Gaussian Process Regression (GPR) that has several advantages over the standard analytical approach. It can model complex non-linear relationships, which perfectly fits within the scope of this problem. GPR provides not only predictions but also estimates of uncertainty, offering valuable insights into the reliability of its forecasts. Another advantage is that it performs well with limited data; as in the case of crash testing, data collection is quite expensive and it is difficult to cover an entire range of vehicles.

https://doi.org/10.14669/AM/189172

2. Data collection

The mathematical model is based on data shared by the National Highway Traffic Safety Administration from the United States. The NHTSA performs crash tests to evaluate the safety of vehicles and to provide consumers with information about their crashworthiness. Three types of tests are performed:

- Frontal Crash Tests: These tests simulate a head-on collision between two vehicles of the same weight and size. Vehicles are crashed into a rigid barrier at a specified speed.
- Side Impact Tests: Simulating a side-impact collision, these tests involve a moving barrier that impacts the side of a stationary vehicle. There are both driver-side and passenger-side tests.
- Rollover Resistance Tests: These tests assess a vehicle's propensity to roll over in a single-vehicle crash. They involve measuring the height and shape of a vehicle and using mathematical models to predict its rollover risk.

Crash tests are conducted using specialised equipment including crash test dummies equipped with sensors to measure forces exerted on the body as well as high speed cameras recording the crash from different angles in order to analyse vehicle structure and other recording instruments capturing crash forces, vehicle movement and structural integrity.

After the crash tests, engineers and researchers analyse the collected data to assess the performance of the vehicle in terms of occupant protection, structural integrity and over-all safety. Results are then used to assign safety ratings to vehicles, which are made available to the public through programs like the New Car Assessment Program (NCAP).

3. Gaussian Process Regression

A regression problem is a common situation in which we would like to determine the value of a given quantity y based on known predictive factors x. This relationship between x and y is usually assumed to be the possibility to represent y as a function of x, the shape of which is usually unknown. While in some cases assuming the linear kind of relationship between inputs x and outputs y is sufficient, this is not the case when speaking of such nonlinear problems as precrash velocity prediction. To make matters worse, a high percentage of nonlinear approaches require rather copious amounts of data, which, in the case of the problem at hand, can be both tedious and costly. Therefore, Gaussian Process Regression (GPR) is the tool we will refer to when dealing with this issue.

Gaussian Process Regression is a non-parametric Bayesian approach, able to work without making any prior assumptions about the type of distribution assumed by the data [4, 9, 10]. This family of models have a very large capacity, being able to learn very complicated

https://doi.org/10.14669/AM/189172

patterns due to not having any intrinsic or imposed bounds on the number of parameters used. The complexity of the resulting mapping between input and output is thus inferred from the data itself through Bayesian inference.

Suppose that there exists an unknown functional relationship between the inputs (in this case, the vehicle mass, the length of the crash surface and the indents), which we will denote with x, and the outputs (precrash velocity), denoted by y. Our task is to obtain the best possible approximation of $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$, $f(x) = y$ for all values of x in the training set, while still being able to generalise this mapping to the previously unseen values of x. Since measurements which were done to prepare the training data are subject to errors, in most cases, we assume that these errors follow the normal distribution $N(0, \theta_\varepsilon^2)$ for some $\theta_\varepsilon > 0$. Thus, we assume that:

$$y = f(\boldsymbol{x}) + \varepsilon \tag{1}$$

where ε is the random variable depicting the errors, which appear no matter how many measurements we take. Unlike in the most Bayesian frameworks, we will also assume that $f(\boldsymbol{x})$, which in the vocabulary of GPR is termed as "signal", is also a random variable, independent from ε and with its own distinctive distribution. In general, we will assume that $f(\boldsymbol{x})$ is a Gaussian process (hence the name of the method) and is thus completely characterised by its mean and covariance functions, which we will denote by μ and k, respectively. The mean function μ is defined as the one satisfying $\mu(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})]$, i.e. it gives us the average value of $f(\boldsymbol{x})$ for a given input $\boldsymbol{x}$. To simplify the computations, the data is usually standardised to achieve $\mu \equiv 0$. This enables us to perform inferencing solely based on the second function, which is the covariance function, also known as the **kernel** function. Its purpose is to model the dependence between the function values at distinct inputs in the following manner:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}[(f(\boldsymbol{x}) - \mu(\boldsymbol{x}))(f(\boldsymbol{x}') - \mu(\boldsymbol{x}'))\,] \tag{2}$$

The selection of the kernel should take into account factors such as smoothness of the data, expected shape of the relationship between inputs and outputs, patterns which may appear in the data etc. If the dataset is sufficiently small, this choice can be dictated by data by trial-and-error procedure, grid search or similar techniques.

In general, the reasonable assumption regarding the kernel should be as follows – the closer two inputs lay, the greater the correlation between the outputs they produce (although assuming that the reverse implication holds would be a critical mistake). The RBF (radial basis function) kernel not only complies with said requirement but is also a rather expressive function for modelling many smooth relationships. The radial basis function kernel is given by the formula:

$$k_{RBF}(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 e^{-\frac{1}{2\lambda^2} \cdot d(\boldsymbol{x}, \boldsymbol{x}')^2} \tag{3}$$

https://doi.org/10.14669/AM/189172

where d denotes the distance between the inputs $\boldsymbol{x}$ and $\boldsymbol{x'}$, while λ and σ are hyperparameters of this kernel.

The length-scale parameter λ determines the length of the highly nonlinear fragments of the approximated function. In general, extrapolation beyond the scope of the interval covered with data points extended by the squared value of λ via GPR becomes highly unstable and hard to rely on.

The output/signal variance σ^2 determines the average distance of the model prediction from the mean. This parameter is common amongst most of the kernels, serving as a scale factor.

Although these hyperparameters can be found via the Bayesian approach [3], this is rarely done in practice due to it being difficult, especially when considering more complex kernels. Instead, we opt for maximum likelihood estimation [9] or grid search [7] as more time-efficient alternatives (although it should be noted that they might be less stable than Bayesian inference).

Once we have decided on the choice of the kernel function, we can use the Gaussian process to draw a priori values along with the posterior function values conditional upon previous observations.

Despite Gaussian processes being continuous, when we sample a function from a Gaussian process, we do it by computing its values on a selected set of inputs. This is usually done by drawing outputs for these points by the means of a multivariate normal distribution with a covariance matrix generated by the kernel in the following manner – we first collect a vector of input points $X = (x_1, \ldots, x_n)$ and compute the covariance matrix as:

$$K(X,X) = \left[k\left(x_i, x_j\right)\right]_{1 \le i,j \le n} \tag{4}$$

Thus, we sample the distribution $N(\mathbf{0}, K(X,X))$, where $\mathbf{0}$ stands for a mean function which equals 0 everywhere – this can be done by applying the previously mentioned simplification. With the obtained vector $\boldsymbol{f}_X$, we can turn it into an observation vector by adding sampled error terms.

Thus far, we have not incorporated the data used to train the regressor in these considerations. We shall do it now – suppose that we have at our disposal a set of pairs of points $\{(x_t, y_t) : \ t \in T\}$. Let us denote by X_T the set of inputs from this dataset and by Y_T the respective outputs. If we'd like to make new predictions on the new points, we should create a distribution based on the previous observations. Thus, we will assume that:

$$\begin{bmatrix} Y_T \\ \boldsymbol{f}_X \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K(X_T, X_T) + \theta_\varepsilon^2 \cdot \boldsymbol{I}_T & K(X_T, X) \\ K(X, X_T) & K(X,X) \end{bmatrix}\right) \tag{5}$$

https://doi.org/10.14669/AM/189172

where $K(A,B)$ is defined as in (KER) and X stands for a vector of new inputs. Additionally, $\boldsymbol{I}$ in the formula above stands for identity matrix of size $|T|$. Since we know both new inputs and the previous dataset, we can derive the conditional distribution of $\boldsymbol{f}_{\boldsymbol{X}}$ based on this knowledge, thus obtaining:

$$\boldsymbol{f}_{\boldsymbol{X}}|X, X_T, Y_T \sim N(\boldsymbol{\mu}, \Sigma) \tag{6}$$

where: $\boldsymbol{\mu} = K(X, X_T) \cdot [K(X_T, X_T) + \theta_\varepsilon^2 \boldsymbol{I}]^{-1} Y_T$,

$$\Sigma = K(X, X) - K(X, X_T) \cdot [K(X_T, X_T) + \theta_\varepsilon^2 \boldsymbol{I}]^{-1} Y_T \cdot K(X_T, X).$$

It is a noteworthy fact, that the posterior distribution of $\boldsymbol{f}_{\boldsymbol{X}}$ can also be perceived as the Gaussian Process, but this does not matter in our considerations. Therefore, predicting $\boldsymbol{f}_{\boldsymbol{X}}$ can be done via taking the means $\boldsymbol{\mu}$ or sampling from said Gaussian Process describing the posterior $\boldsymbol{f}_{\boldsymbol{X}}|X, X_T, Y_T$.

4. Results

The best obtained model utilizing RBF and white noise kernels. Its parameters are summarised in Table 1. Figures 1 and 2 show approximation error in absolute and percentage value respectively.

Tab. 1. Parameters of the best obtained model utilizing RBF and white noise kernels

Parameter name	Value (approximately)
Length-scale l:	10.29580
Signal variance σ^2:	218.49159
White noise variance θ^2:	4.67163

https://doi.org/10.14669/AM/189172

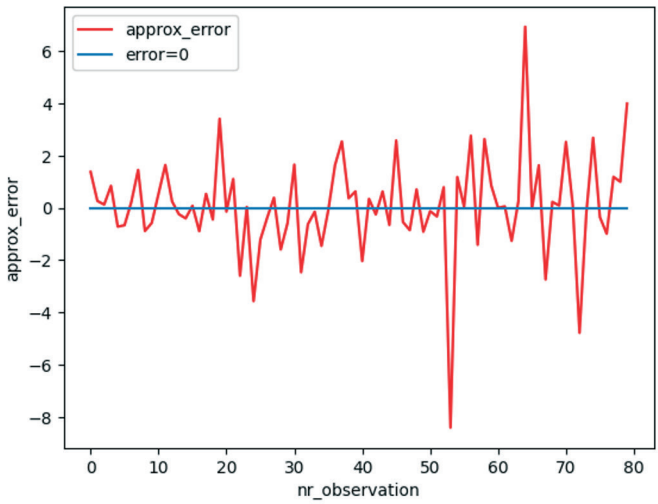

Fig. 1. Approximation error vs observation ID on test data

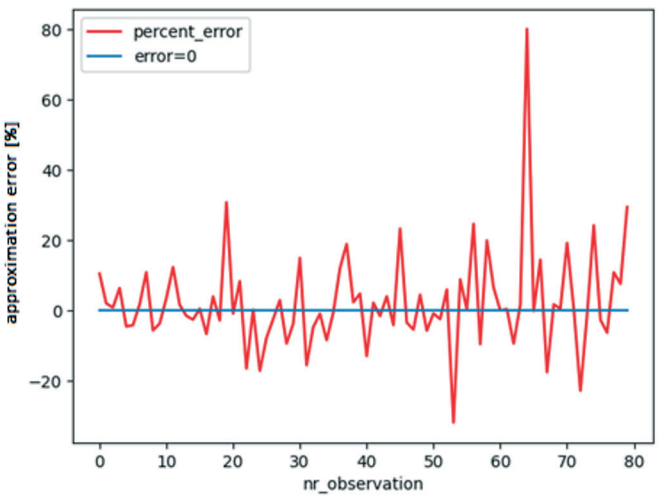

Fig. 2. Percent approximation error vs observation ID on test data

The resulting model scores mean absolute error of magnitude 1.22 m/s (on the test dataset, which consists of 20% of the whole NHTSA dataset for the compact car class), which corresponds to a mean absolute percentage error of 8.30%. The observant reader would surely notice why the percentage error is not a reliant measure of the regression by taking a closer look at both the error plots and their extreme values – heavy underestimation by over 8m/s is equivalent to a less than 40% percentage error, while smaller overestimation of around 7m/s yields over 75% of percentage error. While most of the predictions are within the 2m/s error margin, some outliers to this rule are visible in the approximation error plot.

https://doi.org/10.14669/AM/189172

To allow the reader to more easily compare between predicted and observed data points, we present predictions (along with the uncertainty margins) vs actual observations in Figure 3.

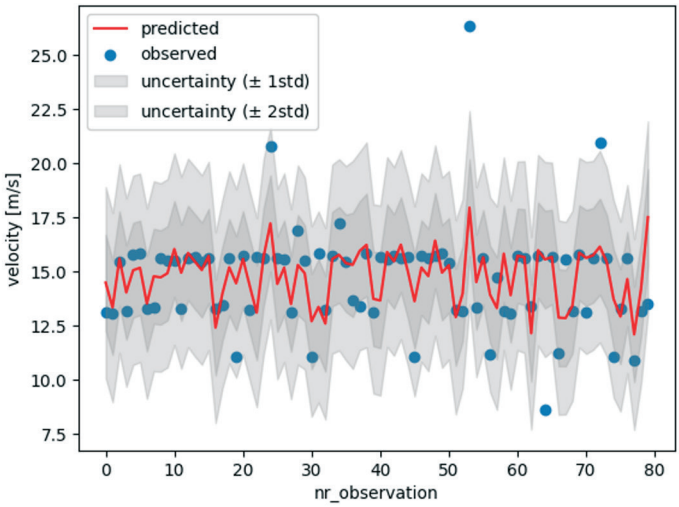

Fig. 3. Predicted vs observed test data

5. Conclusions

A Gaussian Process Regression model for precrash velocity estimation based on the damage done to the car as well as its mass and width was generated. The proposed methodology objectively yields better results than methods based on linear regression. The performance of the model is beyond satisfactory. It does not only provide better estimates of precrash speed but also supplies the user with the uncertainty margins, leading to a clearer picture for both scientific and civil investigators.

In the future, the authors plan to work on a dedicated handheld device for velocity determination. Such a device would scan the deformed portion of the vehicle by means of a ranging laser or photogrammetry, measuring the Cs coefficient and instantly producing the precrash speed.

References

[1] Aleksandrowicz P, Aleksandrowicz I, Kukiełka K, Patyk R, Stanowski P. Problem with determining the vehicle impact velocity for car bodies breaking apart. Transport Problems. 2022;17(3):75–86. https://doi.org/10.20858/tp.2022.17.3.07.

[2] Dudziak A, Caban J, Stopka O, Stoma M, Sejkorová M, Stopková M. Vehicle Market Analysis of Drivers' Preferences in Terms of the Propulsion Systems: The Czech Case Study. Energies. 2023;16(5):2418. https://doi.org/10.3390/en16052418.

[3] Flaxman S, Gelman A, Neill D, Smola A, Vehtari A, Wilson AG. Fast hierarchical Gaussian processes. 2015.

[4] Gershman SJ, Blei DM. A tutorial on Bayesian nonparametric models. Journal of Mathematical Psychology. 2023;56(1):1–12. https://doi.org/10.1016/j.jmp.2011.08.004.

[5] Guan J, Yao Y, Zhao W, Hagiwara I, Zhao X. Development of an Impact Energy Absorption Structure by an Arc Shape Stroke Origami Type Hydraulic Damper. Shock and Vibration. 2023;2023:1–11. https://doi.org/10.1155/2023/4578613.

[6] Guzek M, Jackowski J, Jurecki RS, Szumska EM, Zdanowicz P, Żmuda M. Electric Vehicles—An Overview of Current Issues—Part 2—Infrastructure and Road Safety. Energies. 2024;17(2):495. https://doi.org/10.3390/en17020495.

[7] Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS. preprint. https://doi.org/10.48550/arXiv.1912.06059.

[8] Moravcová P, Bucsuházy K, Zůvala R, Semela M, Bradáč A. What should I use to calculate vehicle EES?. Plos one. 2024;19(2):e0297940. https://doi.org/10.1371/journal.pone.0297940.

[9] Williams CK, Rasmussen CE. Gaussian processes for machine learning Cambridge. MA: MIT Press. 2006.

[10] Williams CK. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. Learning in graphical models. Dordrecht: Springer Netherlands. 1998;89:599–621. https://doi.org/10.1007/978-94-011-5014-9_23.

[11] Zou T, Liu Y, Zhang Y. A method for analyzing accident reconstruction results under complex uncertain conditions. International journal of crashworthiness. 2023;28(2):224–234. https://doi.org/10.1080/13588265.2022.2074721.