Heinrich C. Mayr, Martin Pinzger (Hrsg.): INFORMATIK 2016,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2016 385

Weighted Aggregation in the Domain of Crowd-Based Road Condition Monitoring

Kevin Laubis[1], Viliam Simko[2], Christof Weinhardt[3]

Abstract: This paper focuses on crowd-based road condition monitoring using smart devices, such as smartphones and evaluates different strategies for aggregating multiple measurements (arithmetic mean and weighted means using R^2 and *RMSE*) for predicting the longitudinal road roughness. The results confirm that aggregating predictions from single drives leads to a higher model performance. This has been expected and confirms the intuition. The overall R^2 could be increased from 0.69 to 0.75 on average and the *NRMSE* could be decreased from 9 % to 8 % on average. However, contrary to the intuition, the results show that weighted aggregations of single predictions should be avoided, which is consistent with previous findings in other domains, such as financial forecasting.

Keywords: Crowd-based sensing, road condition monitoring, international roughness index, predictive road maintenance, weighted aggregation, ensemble learning

1 Introduction

Road roughness is one of the most important attributes that gives valuable insights into road condition and driving comfort. Thus, metrics such as the International Roughness Index (IRI) [SGQ86] are considered in most pavement management systems for planning cost efficient road maintenance actions. Nowadays, road authorities relay on measurements from special-purpose vehicles equipped with lasers and further highly precise sensors for sensing the road's profile. Furthermore, specially-trained technicians are required for performing these measurements. A bottleneck in assets and human resources leads to measurements performed on a coarse granular temporal basis. In case of the German's federal road network, this results in monitoring intervals of four years. With regard to road maintenance, this leads to a reactive approach, which directs resources towards road segments that already reached a critical condition.

Using smart devices from drivers and passengers, it is possible to measure and analyze vehicle's vibration and thus, to estimate the road's roughness magnitude. This is a promising alternative to the current way of monitoring the road condition. The inertial measurement units (IMUs) in smart devices allows for a near-real-time assessment of road condition. However, the low accuracy of such sensors, versatile suspension systems, different

[1] FZI Research Center for Information Technology, Information Process Engineering (IPE), Haid-und-Neu-Straße 10–14, 76131 Karlsruhe, laubis@fzi.de
[2] FZI Research Center for Information Technology, Information Process Engineering (IPE), Haid-und-Neu-Straße 10–14, 76131 Karlsruhe, simko@fzi.de
[3] Karlsruhe Institute of Technology (KIT), Institute of Information Systems and Marketing (IISM), Fritz-Erler-Straße 23, 76131 Karlsruhe, weinhardt@kit.edu

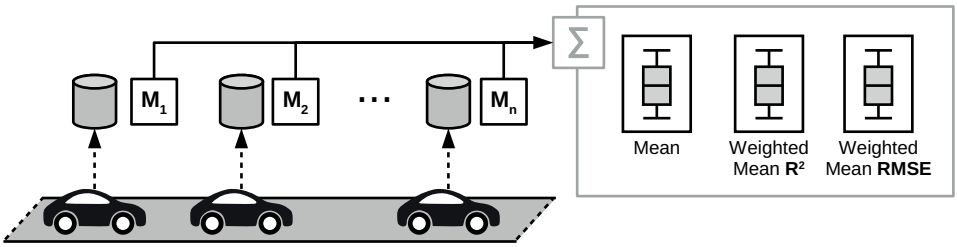

Fig. 1: Outline of aggregating crowd-based road condition measurements from multiple cars

placements of smart devices in the car and other factors lead to a lower prediction accuracy compared to well-calibrated laser-based measurements. For overcoming this decrease in accuracy, multiple approaches for aggregating measurements from several cars are compared in this paper. The outline of this idea is depicted in Fig. 1.

Assuming uncorrelated prediction errors of the single car's predictions, the unweighted mean is expected to reduce the variance component of the errors and thus, increase the prediction accuracy. However, it is not clear to what extent the accuracy can be increased. While it might seem plausible to use weighting schemas based on the model performance when aggregating the results instead of applying a simple arithmetic mean, it has been shown that the arithmetical mean often outperforms a more sophisticated weighting [SW09]. This is true, since a weighted aggregation could increase the prediction error because of increasing the variance component of the error. On the other hand, weighting could reduce the bias component of the prediction error. Thus, it has to be investigated empirically, how weighted aggregation functions perform against the unweighted mean.

For determining, to what extent aggregation of single car predictions can increase the model's accuracy and whether weighting of the single car's prediction is worth an implementation, the focus of this work is to develop and evaluate the extent to increase the performance of crowd-based road roughness estimations by aggregating estimations from multiple cars. We apply unweighted and weighted aggregation methods to a dataset gathered by multiple car drives and laser-based road profile measurements as a ground truth to answer the following research questions:

1. To what extent does the aggregation of crowd-based road roughness measurements from multiple cars increase the model performance?

2. How does the application of weighted aggregation methods instead of unweighted mean affect the model performance?

The paper is structured as follows: The next section 2 summarizes related work. Section 3 describes the data gathered and analyzed from a single car point-of-view followed by the sections 4 and 5, which compare the results from three different aggregation methods. Finally, a conclusion is provided in section 6 and future work is discussed in section 7.

2 Related Work

Investigating road conditions with smart wearable devices or other devices attached to the vehicle, which are able to determine the vehicle's vibrations was goal of several studies [Bh12, Du14, Er08, MPR08, Ni14, Pe11]. They mainly rely on machine learning approaches and on physical models representing a car together with its suspension system ([Ja14]).

The authors of the study [Ni14] show a way of applying supervised machine learning algorithms for predicting the road roughness based on the vertical acceleration provided by high precision sensors. They built models for different road roughness metrics, but did not aggregate multiple measurements.

In [MPR08], the fact that the smart devices could be placed at different locations in the car and with different orientations is considered in addition. This is handled by applying a virtual reorientation of the device's axes. Nevertheless, predictions from single cars were just considered separately.

An approach for determining road segments with single anomalies, such as potholes and bumps was developed in [Pe11]. The authors also applied machine learning algorithms and indicated the single car's performances. An aggregation of multiple measurements was not performed.

A prominent paper in this field is [Er08]. It describes a machine learning approach for also detecting potholes with a fleet of smartphone-equipped taxis. For getting robust results, the pothole candidates from single cars were geo-spatially clustered. However, the performance increase by applying this aggregation function was not investigated.

In none of these studies, the effect of considering measurements from multiple vehicles was investigated. Some of them considered multiple measurements in terms of geo-spatial clustering of singular road anomalies [Er08], but the effect on the overall model performance compared to single measurements was not examined. Since the effect of multiple measurements was not determined yet, likewise the performance of different aggregation functions was not determined in the road condition monitoring domain.

3 Single Car Models

For applying and evaluating different aggregation functions, a set of single-car road roughness prediction models[4] is considered in this study. Each model, predicts the IRI of road segments based on the built-in IMU sensors in a smartphone attached to the dashboard of a car.

The models consider speed of the car as measured by the GPS module ($\sim$ 1 Hz), the 3-axis accelerometer ($\sim$ 50 Hz) and the 3-axis gyroscope sensors ($\sim$ 50 Hz) of the smartphones. The minimum, maximum, standard deviation, variance and the root mean square

[4] Prediction models based on random forests [Br01] are considered in this study.

sensor	aggregation	number of features
GPS velocity	min., max., SD, var., RMS	5
accelerometer (3-axis)	min., max., SD, var., RMS, CWT for 5 bands	30
gyroscope (3-axis)	min., max., SD, var., RMS, CWT for 5 bands	30

Tab. 1: Features extracted from the smartphone's IMU sensors

of each 100 m road segment are extracted and considered as features of the prediction model. From the acceleration and gyroscope measurements, the continuous wavelet transformation (CWT) [TC98] is computed using the *biwavelet* R package [GGS16]. From the bias-corrected wavelet power spectrum, the wavelengths of 5.5 m, 31 m, 124 m, 351 m and 703 m are selected as additional features. As shown in Tab. 1 this results in a set of 65 features in total.

The car models are trained and evaluated on a 4.8 km road link for which the actual road profile is provided by the Institute of Highway and Railroad Engineering at the Karlsruhe Institute of Technology (ISE / KIT). Seven independent car drives are performed and their measurements are matched to the ground truth road roughness for training and testing the corresponding seven prediction models. The models are tuned separately for increasing the coefficient of determination R^2. Next to this performance metric, the root mean square error (*RMSE*) and its normalization to the spread of all IRI values measured in this study (*NRMSE*) are determined. The *NRMSE* is defined as follows:

Def. 1 (Normalized root mean square error)**.** *Let $RMSE \in \mathbb{R}^+$ be the root mean square error of a prediction model. Let $y_{max} \in \mathbb{R}^+$ and $y_{min} \in \mathbb{R}^+$ be the maximum and minimum of the actual laser-based measured IRI values of the considered road link.*

The normalization of the RMSE is defined as follows:

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}}$$

The *NRMSE* was chosen in addition to the *RMSE* for better interpretability of the results with regard to its implications. Since both values are proportional, a particular weighting leads to the same performance of the aggregation functions. Thus, the *RMSE* is chosen as the weight in section 4 and the *NRMSE* values are considered for discussing the results in section 5.

The performance metrics of the single car models considered in this study are shown in Tab. 2. The last four rows provide summary statistics of the performance metrics for an easy comparison with the performances of the aggregated methods described in the following section.

drive	R^2	*RMSE*	*NRMSE*
1	0.6296	0.2063	0.0975
2	0.5947	0.2184	0.1022
3	0.6604	0.2089	0.0977
4	0.7607	0.1697	0.0794
5	0.7899	0.1601	0.0749
6	0.6967	0.1931	0.0903
7	0.7240	0.1924	0.0900
max.	0.7899	0.2184	0.1022
mean	0.6937	0.1927	0.0902
median	0.6967	0.1931	0.0903
min.	0.5947	0.1601	0.0749

Tab. 2: Out of sample performance of single predictions

4 Aggregation Methods

The aggregation of multiple single car predictions is performed in basically two different ways. The baseline aggregation function is an unweighted arithmetic mean of the single predictions. It is formally defined in the following Def. 2.

Def. 2 (Aggregation by unweighted mean)**.** *Let $M_1,\ldots,M_n$ be prediction models. Each model M_i is a function that maps an n-dimensional feature vector $x \in \mathbb{R}^n$ to a real outcome, i.e. $M_i : \mathbb{R}^n \mapsto \mathbb{R}$.*

The predictions $\forall i = 1,\ldots,n : M_i(x) = \hat{y}_i$ can be combined using a simple arithmetic mean $\overline{M}(x)$ as follows:

$$\overline{M}(x) = \frac{\sum_{i=1}^{n} \hat{y}_i}{n}$$

Def. 3 (Aggregation by weighted mean)**.** *Let $M_1,\ldots,M_n$ be prediction models with corresponding weights $W = \{w_1,\ldots,w_n\}$. Each model M_i is a function that maps an n-dimensional feature vector $x \in \mathbb{R}^n$ to a real outcome, i.e. $M_i : \mathbb{R}^n \mapsto \mathbb{R}$.*

The predictions $\forall i = 1,\ldots,n : M_i(x) = \hat{y}_i$ can be combined using a weighted mean $\overline{M}_W(x)$ as follows:

$$\overline{M}_W(x) = \frac{\sum_{i=1}^{n} \hat{y}_i w_i}{\sum_{i=1}^{n} w_i}$$

Now, as weights, the R^2 and *RMSE* can be used for car models $M_1,\ldots,M_n$ which yields aggregation functions: $\overline{M}_{R^2}$ and $\overline{M}_{RMSE}$. The weighted aggregation functions are applied to determine whether attaching importance to more performant models has a positive effect on the prediction performance of the aggregate.

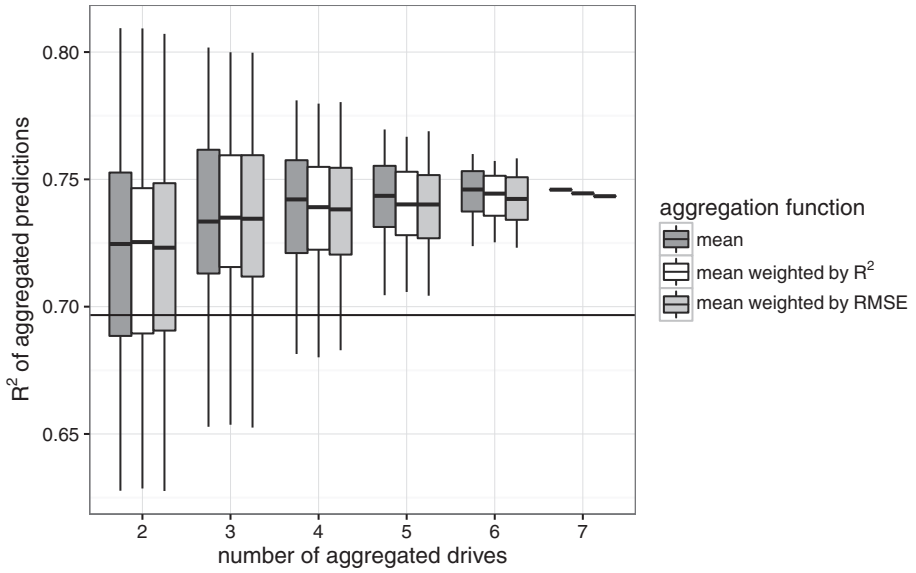

Fig. 2: Out of sample R^2 of aggregated predictions

5 Evaluation

For determining the effect of the aggregation functions and for comparing the performance of the different aggregation functions empirically, they are applied to the models resulting from the drives described in section 3. Since there were $n = 7$ separate drives performed, the combinations from $k = 2, \ldots, 7$ drives are considered. This leads to sample sizes of: $\binom{n}{k} = \{21, 35, 35, 21, 7, 1\}$

The distributions of R^2 over the driving combinations are shown in Fig. 2. For each aggregation function ($\overline{M}$, $\overline{M}_{R^2}$ and $\overline{M}_{RMSE}$) a separate boxplot is shown.

The median R^2 of single drive's predictions (see also Tab. 2) is indicated by a horizontal line. Even for combining two drives, all aggregation functions achieve a significantly better median performance (R^2 and $RMSE$) compared to its single drive baseline at a significance level of at least 5 %. Increasing the number of considered drives likewise increases the coefficient of determination constantly for each aggregation function. Comparing the performance of the aggregation functions shows that there is just a minor difference between them. However, while considering four and more drives, the unweighted mean outperforms the weighted ones.

A comparison regarding the $NRMSE$ is given in Fig. 3. Similar to the R^2 scenario, a constant decrease of the $NRMSE$ is achieved by considering more drives. Furthermore, the unweighted mean aggregation has a lower median $NRMSE$ than the weighted aggregations except for the case of two drives.

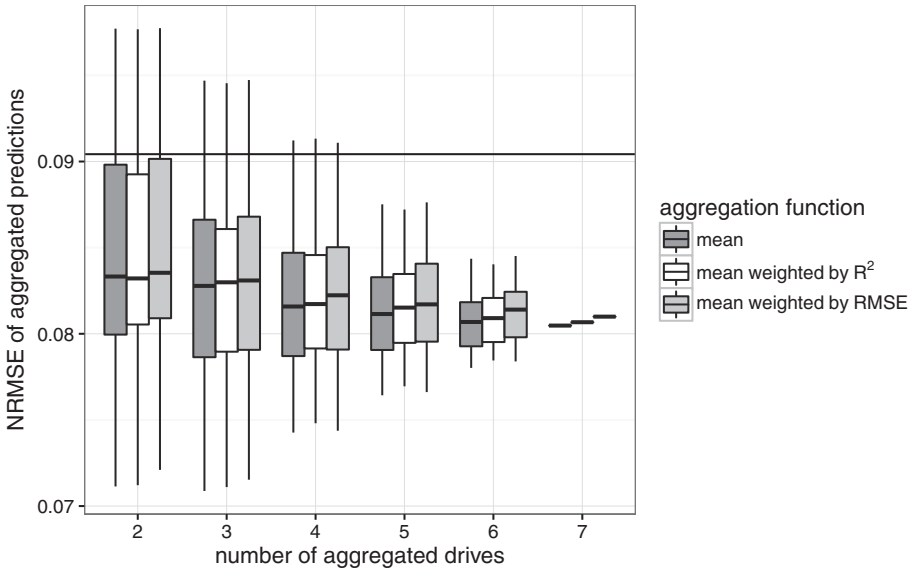

Fig. 3: Out of sample *NRMSE* of aggregated predictions

The mean R^2 and mean *NRMSE* over all aggregation combinations for the considered aggregation functions and for the considered number of drives are given in Tab. 3. If there was a significant performance decrease (decrease in R^2 or increase in *NRMSE*) of using a weighted mean aggregation instead of using the unweighted mean aggregation, it is indicated at the corresponding mean performance of the weighted aggregation. Since the first row indicates the baseline with no aggregation, there is no difference between the aggregation functions. Likewise, referring to the number of possible combinations, the seventh row does not contain tests on significant differences in performance. Except for combinations of drives less or equal to three, the mean performance of the R^2 weighted aggregations are worse than the unweighted aggregations. Regarding the mean performance of the *NRMSE* the unweighted aggregation outperforms the weighted ones for all considered numbers of drives. Even thought the absolute differences are minor, the decreases compared to the unweighted mean function are mostly statistically significant even for the small sample sizes. This indicates that applying a weighted aggregation increases the variance error component to a higher extent than decreasing the bias error component. A vertical comparison of the performance metrics provided in Tab. 3 was discussed based on Fig. 2 and Fig. 3.

6 Conclusion

For answering the research questions (1) to what extent the aggregation of crowd-based road roughness measurements from multiple cars does increase the model performance and (2) how the application of weighted aggregation methods instead of unweighted mean does affect the overall performance, three different aggregation methods (arithmetic mean

	mean		mean weighted by R^2		mean weighted by *RMSE*	
drives	R^2	*NRMSE*	R^2	*NRMSE*	R^2	*NRMSE*
1	0.6937	0.0902	0.6937	0.0902	0.6937	0.0902
2	0.7234	0.0846	0.7227	0.0846	0.7223*	0.0849***
3	0.7338	0.0827	0.7327**	0.0828	0.7320***	0.0831***
4	0.7391	0.0817	0.7379***	0.0819**	0.7369***	0.0822***
5	0.7423	0.0811	0.7409***	0.0813**	0.7399***	0.0816***
6	0.7444	0.0808	0.7430**	0.0809*	0.7419***	0.0813***
7	0.7460	0.0805	0.7445	0.0807	0.7434	0.0810
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$						

Tab. 3: Means of R^2 and *NRMSE* over aggregation combinations for considered aggregation functions (unweighted mean, mean weighted by R^2 and mean weighted by *RMSE*) and for considered number of drives. Significant performance decreases compared to unweighted mean aggregations are indicated.

and weighted means using R^2 and *RMSE*) are applied to seven single car models. Each of the car models is a random forest trained on sensor data collected using a smartphone while driving a distance of $4.8\,km$.

The results confirm that aggregating predictions from single drives leads to a higher model performance. This has been expected and confirms the intuition. Thus, the R^2 could be increased from 0.69 to 0.75 on average and the *NRMSE* could be decreased from $9\,\%$ to $8\,\%$ on average. In other words, real-time predictive road maintenance gets better with increasing number of participants.

Contrary to the intuition, our results also show that weighting aggregations of single predictions should be avoided. This is consistent with the results of the study [SW09], which describes similar findings in the financial forecasting domain. From a technical point-of-view, this allows a simpler and thus, more efficient implementation.

7 Outlook

It has to be mentioned that all measurements are performed on a homogeneous and recently paved road link. Thus, it is not clear whether the results are valid for other road types and for a wider IRI range. Beside extending the analysis by a broader set of road segments and by further car and sensor types, a steps will be to investigate whether the results are reasonable for road conditions other than IRI as well. Furthermore, it is intended to determine how the real-time road condition monitoring affects the road maintenance from a managerial point of view. It has to be investigated to what amount less accurate models could be applied, while still being economically beneficial for road authorities. It should be determined whether single car predictions are already sufficiently accurate for planning maintenance actions and if not, it is going to be determined whether the accuracy increase achieved by aggregation reaches a sufficient accuracy level.

Knowing this economic value of a crowd-based road roughness monitoring for road authorities and for road users a business model can be tailored. Such a model should also encompass incentive mechanisms for motivating drivers to participate in the crowd-based system.

References

[Bh12] Bhoraskar, Ravi; Vankadhara, Nagamanoj; Raman, Bhaskaran; Kulkarni, Purushottam: Wolverine: Traffic and road condition estimation using smartphone sensors. IEEE, pp. 1–6, jan 2012.

[Br01] Breiman, Leo: Random forests. Machine learning, 45(1):5–32, oct 2001.

[Du14] Du, Yuchuan; Liu, Chenglong; Wu, Difei; Jiang, Shengchuan: Measurement of International Roughness Index by Using Z-Axis Accelerometers and GPS. Mathematical Problems in Engineering, 2014:1–10, 2014.

[Er08] Eriksson, Jakob; Girod, Lewis; Hull, Bret; Newton, Ryan; Madden, Samuel; Balakrishnan, Hari: The pothole patrol: using a mobile sensor network for road surface monitoring. In: Proceeding of the 6th international conference on Mobile systems, applications, and services - MobiSys '08. ACM Press, New York, New York, USA, p. 29, 2008.

[GGS16] Gouhier, Tarik; Grinsted, Aslak; Simko, Viliam: . R package "biwavelet": Conduct univariate and bivariate wavelet analyses, 2016. (Version 0.20.4).

[Ja14] Jazar, Reza N.: Vehicle Dynamics. Springer New York, New York, NY, 2 edition, 2014.

[MPR08] Mohan, Prashanth; Padmanabhan, Venkata N.; Ramjee, Ramachandran: Nericell: Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones. In: Proceedings of the 6th ACM conference on Embedded network sensor systems - SenSys '08. ACM Press, New York, New York, USA, p. 323, 2008.

[Ni14] Nitsche, P.; Van Geem, C.; Stütz, R.; Mocanu, I.; Sjögren, L: Monitoring ride quality on roads with existing sensors in passenger cars. In: Proceedings of the 26th ARRB Conference. Sydney, pp. 1–13, 2014.

[Pe11] Perttunen, Mikko; Mazhelis, Oleksiy; Cong, Fengyu; Kauppila, Mikko; Leppänen, Teemu; Kantola, Jouni; Collin, Jussi; Pirttikangas, Susanna; Haverinen, Janne; Ristaniemi, Tapani; Riekki, Jukka: Distributed Road Surface Condition Monitoring Using Mobile Phones. In: Lecture Notes in Computer Science, pp. 64–78. 2011.

[SGQ86] Sayers, Michael W.; Gillespie, Thomas D.; Queiroz, Cesar A V.: The International Road Roughness Experiment – Establishing Correlation and a Calibration Standard for Measurements. Technical report, Washington, D.C., 1986.

[SW09] Smith, Jeremy; Wallis, Kenneth F.: A Simple Explanation of the Forecast Combination Puzzle. Oxford Bulletin of Economics and Statistics, 71(3):331–355, jun 2009.

[TC98] Torrence, Christopher; Compo, Gilbert P.: A practical guide to wavelet analysis. Bulletin of the American Meteorological society, 79(1):61–78, jan 1998.