<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *IEEE Symposium on Computers and Communications (ISCC) 29 june-3 july 2019, Barcelona, Spain*.

Citation for the original published paper:

Garcia, J., Sundberg, S., Brunstrom, A. (2019)

LTE for Trains - Performance Interactions Examined with DL, ML and Resampling

In: *2019 IEEE Symposium on Computers and Communications (ISCC)* (pp. 1-6). IEEE

<https://doi.org/10.1109/ISCC47284.2019.8969727>

N.B. When citing this work, cite the original published paper.

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kau:diva-76644>

LTE for Trains - Performance Interactions Examined with DL, ML and Resampling

Johan Garcia, Simon Sundberg, Anna Brunstrom

Department of Mathematics and Computer Science

Karlstad University, Karlstad, Sweden

Email: {johan.garcia, simon.sundberg, anna.brunstrom}@kau.se

Abstract—Current LTE networks provide a large fraction of the mobile communication needs. One recent application area that have attained additional interest is the provision of mobile communication services to train passengers. To allow more efficient use of network resources and better onboard communication experience, onboard traffic aggregation can be performed. In this work we examine a large-scale operational data set from a router-based LTE traffic aggregation system mounted onboard more than 100 trains belonging to a major Swedish train operator. We use both deep learning (DL) with Deep Neural Networks and traditional machine learning (ML) with Random Forests to examine an observed association between train velocity and achieved throughput, which curiously varies over different radio conditions. More than 37000 train journeys are analyzed to explore for structure and learn potential explanatory features. The results indicate that the association has a limited presence on a per cell basis, and that there is only a limited amount of learnable structure per cell. A resampling evaluation shows that the association becomes apparent when cell measurements are aggregated at an order of tens to a hundred cells.

I. INTRODUCTION

Train travel is increasingly becoming the preferred mode of transportation between many cities. During a train journey, the ability to access the Internet provides a considerable utility for most passengers. Internet connectivity can be provided directly by cellular networks such as LTE. However, the signal path loss between an user equipment (UE) inside the carriage and the eNodeB base station is considerable. There are numerous challenges to providing cellular data access to large numbers of user onboard trains as surveyed in [14], [16]. While several approaches have been suggested for providing Internet access onboard trains [13], [9], this work focuses on the case where an onboard router aggregates traffic from many users onboard a train. The particular system examined here provides WiFi access inside the train carriages and a specialized router aggregates the traffic and transfers it over multiple cellular connections using rooftop antennas. Here, we examine the operational data from a large-scale deployment of such a system.

Many factors influence the achieved cellular link throughput, but specifically for a train use case, previous work have indicated that there is a unexpected association between achieved link throughput and train

velocity. In [10], the association between train velocity and achieved throughput was found to be switching between negative to positive as radio conditions improved. An examination of the relationship between velocity and throughput at different SINR suggested sigmoid-like behavior where the effect changed from negative to positive at a point around 12dB. The analysis in [10] was primarily carried out from a per train vantage point using a smaller data set. A later study [11] made an extensive study over a three year period, considering additional vantage points. Results from [11] indicated that the velocity and throughput association was not present when data from individual cells were analyzed. In the current study we extend the per cell analysis to cover all relevant cells in the data set, and examine the cells for any systematic structure coupled to the presence or absence of a sigmoid-like behavior for a cell. We use deep neural networks and random forests to attempt to learn structure, with results indicating that there is only a moderate amount of structure that can be learnt from the per cell data. A resampling analysis showed that data with contributions from a larger number of different cells also showed more sigmoid-like structure.

II. RELATED WORK

There are a number of factors that may influence the capacity observed over a LTE ENodeB-UE link. Radio propagation aspects, cell load, transient environmental conditions, and UE capabilities are factors that may have influence on the obtained link capacity. Fernandez et al. [7] present a survey of environmental effects particular to wireless communication in a railway setting. Using a test train to perform trial runs towards an eNodeB under their control, Dominguez et al. [5] performs an active characterization of LTE characteristics as observed by the train. An interaction effect between train velocity and MIMO rank was first noticed in early LTE train measurements reported by Alasali et al. [1]. Later work by Beckman et al. [2] considers MIMO and propagation as possible factors for a train velocity and link throughput association.

An initial study on a 6 month subset of the current data set was performed by Garcia et al. [10]. In that study, an initial analysis of the train velocity and link throughput

association was performed. The analysis was done for three different train lines, with a focus on grouping the data across individual trains that fulfilled a requirement of a minimum number of observations (>10000 observations per train). A recent followup study [11] utilized a larger data set and examined differences between six train lines and two operators, and also indicated that velocity and throughput association was not present when data was evaluated on a per cell basis. The work reported here utilizes DL and ML to perform a deeper per cell examination, focusing on the four train lines and single operator identified as particularly relevant in [11].

III. DATA COLLECTION

The data used for this evaluation was recorded during 2015-2018. Each train is equipped with an onboard router which uses four Sierra Wireless MC7710 LTE modems, with two modems assigned to each of two cellular operators. Data is collected at five second intervals, and comprises of general metrics such as number of active devices, GPS positions and current velocity of the train, as well as radio-related metrics and performance metrics such as aggregated and per link throughput and ping delay for each of the four links. The router monitors the individual communication conditions on each link and schedules traffic over links according to a proprietary scheme.

Before performing any analysis, the data is adapted and filtered. Spurious zero values are reported for various metrics, possibly due to router reboots, modem failures or other causes. Journeys with such data are filtered out, along with those where values are outside the possible operating conditions.

When using link throughput as a metric an issue arises concerning which factor is limiting the achieved link throughput. As data collection is done in a purely passive way, there are two possible factors that limit the link throughput, either the achievable throughput on the wireless 4G link, or the aggregate traffic load requested by the users. As the analysis here focuses on analyzing the maximum achievable wireless link throughput, only data for the link with the highest throughput is considered in the further analysis.

Train journeys performed along four different stretches of rail track were included in the data set: Stockholm - Göteborg, Stockholm - Malmö, Stockholm - Karlstad, and Stockholm-Umeå. The geographical outline of these train lines are shown in Figure 1. As can be seen, the lines are sharing some sections, but are mostly non-overlapping. Data from over 37000 journeys are included in the data set. The data set contains 61 unique router ids, which can be expected to correspond to the number of unique train sets. Train-line details are provided in Table I.

Figure 1: Train lines used in evaluation

Table I: Examined train lines

Route name	Route Length	Nr of journeys	Nr router ids	Nr of cell-ids	Avg velocity
GbgSto1	485 km	16039	58	2736	143 km/h
MalSto0	614 km	10327	39	3336	136 km/h
StoUme0	709 km	3565	19	3402	109 km/h
KsdSto0	324 km	7378	57	2104	130 km/h
Overall		37309	61	8906	130 km/h

IV. VELOCITY AND THROUGHPUT INTERACTION

The underlying association that is of interest is between train velocity and achieved link throughput. An example of this base association is shown in Figure 2. The figure shows, for one train, a scatter plot of values for train velocity and link throughput, along with a linear regression fitting. While there is a large degree of variation, a negative slope is apparent in the regression which is unlikely to appear due to chance. The data shown in the figure is for a particular subrange of radio conditions, achieved by binning all observations for one train into separate 3 dB wide SINR bins. The data shown in Figure 2 is for the SINR interval -0.5 to 2.5 dB (for observations with a SINR reading, 18% of total). The linear regression results in a v-tp slope coefficient of -4.15 kbps per km/h, meaning that for these relatively poor radio conditions there is a weak association between higher train velocity and lower observed link throughput.

This v-tp slope value results represent the association

Figure 2: Scatter plot between train velocity and link throughput for one train at SINR -0.5 to 2.5 dB.

Figure 3: v-tp slope coefficient over SINR bins

Figure 4: Linear regression of sinr-vtp slope

between link throughput and train velocity for a single train at a single SINR interval. Examining the slope coefficient can also be done over a range of SINR intervals for all trains as illustrated in Figure 3. As can be observed, the relationship between increased train velocity and link throughput has a curious switch from being negative for unfavorable radio conditions, to becoming positive for more favorable radio conditions (i.e as the SINR value increases). In Figure 3 a blue curve is also shown which is a fitted sigmoid curve (see [10] for details).

Figure 4 shows a similarly fitted blue sigmoid curve, but here calculated over all observations without any per train grouping. The resulting sigmoid is very similar showing that the average results are not impacted by per train grouping. As we want to examine a large number of variations of groupings we compute a linear regression approximation to the sigmoid curve for use in the further analysis. The linear approximation is shown as a green line in figure, and the sinr-vtp slope coefficient (0.47) for this line is given in the figure. Further, it can be seen that the green line crosses the y-axis zero line at around 9 dB.

V. PER CID MACHINE LEARNING ANALYSIS

Motivated by the initial results for 48 cell ids (CIDs) reported in [11], we perform a more comprehensive study of the potential presence of sigmoid-like behavior on a per CID basis.

A. Initial statistical examination

We compute the sinr-vtp slope and zero crossing values for each of the 790 CIDs in our data set which have at least 10000 observations. The distribution of these slope and zero crossing values are shown in blue in Figure 5. The figure also shows the distribution when

(a) Distr. of sinr-vtp slope coeffs (b) Distr. of sinr-vtp zero crossings

Figure 5: Distributions of sinr-vtp slopes and zero crossings for data grouped on 790 CIDs or 56 trains

(a) CID grouping

(b) Train grouping

Figure 6: Normal probability plots of sinr-vtp slopes

the observations are grouped according to the trains, i.e. the distribution of slopes and zero crossings resulting from linear fits to each of the trains shown in Figure 3. The per train distribution is shown in apricot, and it is considerably more concentrated than for the CID grouping as well as having a distinctively different slope value. The mean slope value is 0.39 for train and 0.03 for CID groupings. The number of observations per sinr-vtp calculation is different for the two cases. However, tests show that the per train results are not materially different when the number of observations are reduced to match those seen for the per CID grouping.

Further understanding of the data can be obtained by examining the normality of the data. Figure 6 shows normal probability plots [4] for the sinr-vtp slopes for the train grouping and a 95% trim of the CID grouping. Included in the figure is the coefficient of determination, R^2 , for a fitted normal distribution. As can be seen, the CID grouping lies further from normality than the train grouping. To examine if the CID grouped data may have a weak component related to sigmoid-like behavior, a Gaussian Mixture Modeling (GMM) fitting over a range of one to ten components were performed. The best model, as selected by the Bayesian Information Criteria (BIC) [15], is shown in Figure 7. The minor second component did however not exhibit any sigmoid-like tendencies, having a negative μ and very high σ .

This examination of the data on a per CID level did not exhibit any marked sigmoid-like behavior, displaying a much more spread out distribution of both sinr-vtp slope

Figure 7: GMM for CID grouping, with 2 components

Table II: Calculated CID features

eNodeB specific	CID specific	Metric statistics	Binned
n_cids	size	pa	n_bins
n_sectors	n_journeys	ping	bin_size
n_lb	lb	v	bin_delta_v
	lbw	l_tprx	bin_delta_l_tprx
		sinr	slope
		rsrp	zero_crossing
		txp	
		passage_distance	
		passage_duration	

coefficients and zero crossing values which confirms the initial results in [11]. Although no strong sigmoid-behavior were observed, we are interested in examining if there are any systematic characteristics among the CIDs that do fall into the category of exhibiting sigmoid-like behavior. To examine this, we utilize machine learning methods. If there are any structure, i.e. if the CIDs with sigmoid-like behavior have some structural aspects that to some degree distinguish them from those CIDs that do not have sigmoid-like behavior, then this structure will be learnt by the classifier during the training phase assuming features capturing the relevant structural aspects are present. The problem is thus cast as a classification problem, aiming to build a classifier that can predict if a CID employs sigmoid-like behavior or not based on a set of features that capture various attributes of the CIDs and the set of observations of belonging to that CID.

B. Per CID descriptive features

In order to perform machine learning, a set of features are necessary. As the focus here is on identifying structural aspects of CIDs showing sigmoid-like behavior, the focus is on features describing CID aspects. Table II shows an overview of the features used for each CID, grouped into four categories.

The eNodeB specific features capture aspects related to the eNodeB (LTE base station) the CID belongs to. Each eNodeB may have multiple different cells covering different sectors or using different frequency bands. Here, n_lb is the number of unique frequency bands observed for the eNodeB, and n_sectors is the number of observed sectors for the eNodeB.

Features related to a specific CID are lb and lbw, the frequency and the carrier bandwidth used by the cell, and the number of observations belonging to the CID.

A number of metrics are computed derived on the data collected by routers while the link is associated with the CID, as listed in the metric statistics columns. Here, pa is the number of connected devices on the train, l_tprx is the downlink throughput on a single link, and txp is the modem transmission power. The features passage_distance and passage_duration is the distance the train traveled along the rail while connected to the cell and the time it was connected to the cell, respectively. For these metrics the minimum, maximum, mean, standard deviation and variance are computed and used as features.

Finally, the binned values are related to the SINR binning process being performed with the data for each CID. The binned feature n_bins is simply the number of 3 dB wide SINR bins that could be created with at least ten observations in each bin. For bin_size, bin_delta_v and bin_delta_l_tprx, features based on the minimum, maximum and mean value has been calculated for the number of entries in each bin, the difference between the highest and the lowest velocity in each bin, and the difference between the highest and the lowest link throughput in each bin. In total, 62 features are used.

C. Deep learning analysis

Deep learning approaches have shown excellent performance in highly complex classification tasks when adequately designed and parameterized [12]. Here, we employ deep learning in the form of a Deep Neural Network (DNN), to examine if the ability to create more complex, composite features in the inner layers of the DNN would allow structure to be learnt, i.e. achieve high classification performance.

For the experimental evaluation three and four layer neural networks are created with relu activation in the middle layers, and a grid search used to explore different sizes of the layers. The grid search used explores middle layer(s) with sizes of 5-100 neurons. As DNNs are sensitive to the relative magnitude of the input data, scaling is performed to zero mean and unity standard deviation. The evaluation results are shown in Figure 8 as Receiver Operating Characteristics (ROC) curves. ROC curves show all possible tradeoffs between high true positive rate and high false positive rate that the classifier can achieve by varying the decision function. The Area under the ROC curve (ROC-AUC) can be used as a point estimate of classifier performance, and is also given in the figure. A ROC-AUC value above 0.5 represents performance better than random guessing. Here, the DNN (blue curve) achieves a ROC-AUC of 0.64. Such a relatively low value is indicating that only a limited amount of structure could be identified, likely because the underlying data in itself lacks strong structure.

To validate our setup, and contrast the performance when the structure has been purposely removed, we also perform a permutation test. Figure 8 also shows

Figure 8: CNN and RF classification performance

Figure 9: RF Feature Importance

the results when the same data set is used, but the velocity data points in the underlying data set has been randomly permuted. This removes any structure and gives an illustration of the achievable classification performance when there is no structure. As seen, the performance is worse than for the non-permuted case, giving an indication that the non-permuted case picks up actual structure and not random noise. One weakness of deep-learning approaches is that they make it hard to gain understanding of the created model, and other approaches are often used to gain more structural insight.

D. Random forest evaluation

To gain further understanding about the structure of sigmoid-behaving CIDs, we additionally employ the random forest approach. Random Forests (RF) [3] is an ensemble based approach to develop classification and regression models. In [8] RF variations were considered the best out of 199 evaluated classifiers from 17 families. A strength of the RF approach is that it does not require any data normalization, is robust to the presence of non-informative or correlated features and can natively provide a measure of feature relevance, which gives some understanding of the structure of the data. As

seen in Figure 8, RF performs very similar to DNN. This indicates that the DNN approach did not identify any complex composite features beyond what RF can capture. To obtain further insights on the moderate amount of structure that could be discerned by both DNN and RF, the feature importances as reported by the RF classifiers were considered. Figure 9 shows the averaged feature importances over 50 RF instantiations for the twelve highest ranked features, with the error bars showing the standard deviation. Features related to the underlying v-tp regression are markedly present, as well as various metrics indirectly coupled to the amount of observations per CID. Also, variations on the number of activated devices (pa) are present. Results in [11] show that the sigmoid behavior is unlikely to occur for observations with low number of activated devices, so CIDs with low number of activated devices may have slightly less sigmoid like behavior. However, as the amount of structure learnt by the classifier is low, care should be taken not to over-interpret these feature importances.

VI. RESAMPLING-BASED EVALUATION

The preceding analysis suggests that there is only negligible sigmoid-like behavior in the data when it is analyzed on per CID basis. However, there is a pronounced effect when other groupings of the data are used, such as when grouped per train (Figure 3), or when the whole data set is in one group (Figure 4). These other groupings will mix observations from many CIDs into the groups being analyzed. It is relevant to consider how the sigmoid-like behavior of groupings evolves when the groups contain varying numbers of CIDs. To explore this we employ an approach inspired by resampling-based statistical methods [6].

Each resampling run consists of randomly selecting n CIDs from the data set. From the selected CIDs, 20000 observations are randomly drawn, and linear regression is performed to obtain the sinr-vtp slope and zero crossing (ZC) values for the resampling run. For each value of n , 1000 resampling runs are performed. The distribution of the sinr-vtp slope and ZCs are shown in Figure 10. The resampling distribution when $n = 2$ is shown topmost, and largely resembles the results for the case when all observations of each CID are used to compute the slopes and ZCs, as shown in Figure 5. As n increases, the resulting distributions gradually move towards being more concentrated around the sigmoid-like regions. When $n = 500$, the distributions indicate a high degree of sigmoid-like behavior, similar to the per train distribution shown in Figure 5. From these results it can be concluded that the underlying cause for the sigmoid-like behavior likely is not strongly tied to aspects observable at individual CIDs, but rather to factors that come into play when observations from many CIDs are considered.

Figure 10: Resampling from different number of CIDs

VII. CONCLUSIONS

LTE networks are highly complex systems, and the high-speed train use case adds additional challenges. This work examines interaction effects between train velocity and obtained link throughput that has been observed in a large operational data set, with focus on a per cell vantage point. Deep neural networks are used to examine any presence of structural factors linked to the cells that show a consistent radio quality dependent association between train velocity and link throughput.

Only a weak presence of any structure was indicated. Random forests were also employed, leading to practically the same classification performance, and additionally indicating which features have predictive value for the velocity throughput association on an individual CID level. Finally, resampling methods are applied to examine at what number of aggregated CIDs the consistent association shown in earlier studies becomes apparent. These results show a steady increase in consistent association as the number of CIDs increase, and at 100 CIDs the observed results approximate earlier results. Our study also illustrate the potential usefulness of combining resampling and ML methods, such as

when performing a larger study of the underlying factors contributing to the sigmoid-like behavior that is observed on an aggregate level.

ACKNOWLEDGMENTS

The authors wish to thank Claes Beckman and Mats Karlsson for technical discussions, Tobias Vehkajärvi and Peter Eklund for assisting with data collection and processing, Jari Appelgren and Abdullah Almasri for discussions on statistical methods, and the Swedish National Infrastructure for Computing (SNIC) for providing resources at HPC2N for performing some of the computations. Funding for this study was provided by the HITS project grant from the Swedish Knowledge Foundation.

REFERENCES

- [1] M. Alasali, C. Beckman, and M. Karlsson, "Providing internet to trains using mimo in lte networks," in *2014 International Conference on Connected Vehicles and Expo (ICCVE)*, Nov 2014, pp. 810–814.
- [2] C. Beckman, J. Garcia, S. Alfredsson, and A. Brunstrom, "On the impact of velocity on the train-to-earth mimo propagation channel: Statistical observations and qualitative analysis," in *2017 IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting*. IEEE, 2017, pp. 1865–1866.
- [3] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [4] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey, "Graphical methods for data analysis," *Cole Statistics/Probability Series*, Wadsworth, 1983.
- [5] T. Domínguez-Bolaño, J. Rodríguez-Piñero, J. A. García-Naya, and L. Castedo, "Experimental characterization of lte wireless links in high-speed trains," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.
- [6] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.
- [7] N. Fernandez, S. Arrizabalaga, J. Añorga, J. Goya, I. Adín, and J. Mendizabal, "Survey of environmental effects in railway communications," in *Nets4Cars/Nets4Trains/Nets4Aircraft*. Springer, 2018, pp. 56–67.
- [8] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.
- [9] D. T. Fokum and V. S. Frost, "A survey on methods for broadband internet access on trains," *Communications Surveys & Tutorials*, *IEEE*, vol. 12, no. 2, pp. 171–185, 2010.
- [10] J. Garcia, S. Alfredsson, A. Brunstrom, and C. Beckman, "Train velocity and data throughput-a large scale lte cellular measurements study," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. IEEE, 2017, pp. 1–6.
- [11] J. Garcia, S. Sundberg, A. Brunström, and C. Beckman, "Interactions between train velocity and cellular link throughput - an extensive study," in *Under submission*, 2019.
- [12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.
- [13] É. Masson, M. Berbineau, and S. Lefebvre, "Broadband Internet access on board high speed trains, A technological survey," in *Communication Technologies for Vehicles*. Springer, 2015.
- [14] M. K. Müller, M. Taranetz, and M. Rupp, "Providing current and future cellular services to high speed trains," *Communications Magazine*, *IEEE*, vol. 53, no. 10, pp. 96–101, 2015.
- [15] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [16] J. Wu and P. Fan, "A survey on high mobility wireless communications: Challenges, opportunities and solutions," *IEEE Access*, vol. 4, pp. 450–476, 2016.