Traffic Flow Prediction via Convolutional Deep Learning

Stefano Bilotta, Enrico Collini, Paolo Nesi, Gianni Pantaleo
University of Florence, Distributed Systems and Internet Technology, DISIT Lab
https://www.disit.org , Https://www.snap4city.org, corresponding: paolo.nesi@unifi.it

City Vehicle Flow

Real Time traffic flow data can be measured, collected and exploited

congestion detection and reduction

origin destination matrixes

incident management

optimization of public transport

reducing fuel consumption and emission of pollutants

City Vehicle Flow

Real Time traffic flow data can be measured, collected and exploited

traffic information systems

dynamic route guidance

planned routing

road digital signage

infrastructure planning

Short-Term Traffic Flow Prediciton

…up to 1 hour in advance, with a resolution of 10 minutes.

- State-of-the-art best **AI architecture**
- Relevant **features** in prediction computation
- **validation** performed on real-world road infrastructure

RF, XGBOOST, DNN, LSTM, BI-LSTM, Autoencoder BI-LSTM, Attention CONV-LSTM, CONV-BI-LSTM

traffic, datetime, seasonality, temporal, weather, pollutants

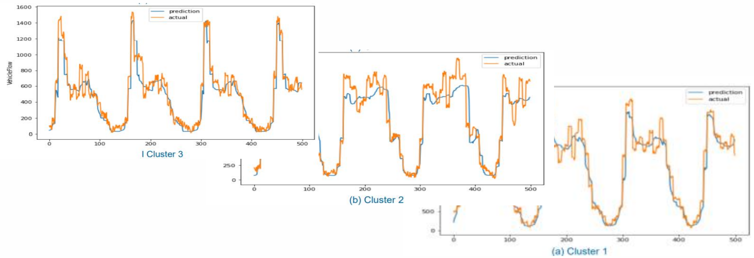
TABLE II -- THE MAPE ESTIMATED FOR 64 COMBINATIONS OF FEATURES FOR ALL THE IDENTIFIED TECHNIQUES AS THE MEDIAN VALUE ON THE SENSORS IN THE 3 CLUSTERS DESCRIBED ABOVE. THE ORDER IS BASED ON THE COMBINATION OF FEATURES. IN BOLD, BEST RESULTS/CONFIGURATIONS. IN BOLD WITH CITATION: RESULTS OBTAINED TAKING INTO ACCOUNT SOLUTIONS FROM THE STATE OF THE ART. PLEASE NOTE THAT CONV-BI-LSTM OVERCOMES ALL OF THEM IN THE SAME FEATURE CONDITIONS.

ID	Features adopted in the model						Median value of MAPE for prediction results by technique								min
	Date time	Traf plus	Temp oral	Season ality	Airpoll	weath er	RF	XGBO OST	DNN	LSTM	BI-LSTM	Autoencod er BI-LSTM	Attention CONV-LSTM	CONV-BI-LSTM	
C1	Y	Y	Y	Y	Y	Y	29.342	34.552	42.754	49.407	34.865	34,708	37,059	31.365	29.342
C2	Y	Y	Y	Y	Y	N	29.682	35.545	43.400	49.832	35.870	35,707	39,506	35.613	29.682
C3	Y	Y	Y	Y	N	Y	28.782	34.441	35.465	36.824	31.555	32,998	33,179	30.894	28.782
C4	Y	Y	Y	Y	N	N	30.935	35.373	38.942	35.383	30.564	32,969	35,713	32.485	30.564
C5	Y	Y	Y	N	Y	Y	29.776	34.469	33.425	42.301	39.865	37,167	35,161	36.897	29.776
C6	**Y**	**Y**	**Y**	**N**	**Y**	**N**	**29.598**	**35.547**	**33.865**	**36.792**	**35.097**	35,322	29,923	**25.981**	**25.981**
C7	Y	Y	Y	N	N	Y	29.421	33.711	31.377	34.736	40.510	37,110	30,741	30.106	29.421
C8	Y	Y	Y	N	N	N	31.245	34.414	32.026	37.823	40.662	37,538	31,263	30.500	30.500
C9	Y	Y	N	Y	Y	Y	29.626	36.919	42.187	**37.068 [3]**	34.297	35,608	36,651	**31.115**	29.626
C10	Y	Y	N	Y	Y	N	29.964	35.802	47.201	41.334	34.743	35,272	40,658	34.116	29.964
C14	Y	Y	N	N	Y	N	29.764	36.374	36.203	43.510	35.744	36,059	33,015	29.827	29.764
C15	Y	Y	N	N	N	Y	29.972	35.423	31.526	46.201	37.209	36,316	32,919	34.313	29.972
C16	Y	Y	N	N	N	N	**30.960 [1]**	34.235	30.338	**37.068 [2]**	**38.082 [4]**	**34,235[5]**	**29,455[6]**	**28.573**	28.573
C17	Y	N	Y	Y	Y	Y	29.281	34.503	72.909	64.557	48.685	41,594	51,026	29.144	29.144
C18	Y	N	Y	Y	Y	N	30.184	35.350	59.458	68.127	46.874	41,112	44,810	30.163	30.163
C27	Y	N	N	Y	N	Y	28.986	35.218	57.938	50.333	59.419	47,318	43,298	28.658	28.658
C28	**Y**	**N**	**N**	**Y**	**N**	**N**	**31.068**	**35.878**	**66.634**	**50.957**	**55.096**	45,487	47,097	**27.561**	**27.561**
C29	Y	N	N	N	Y	Y	29.301	37.532	38.325	40.677	50.303	43,917	35,554	32.784	29.301
C30	Y	N	N	N	Y	N	29.323	37.284	37.149	48.801	55.064	46,174	34,721	32.294	29.323
C31	Y	N	N	N	N	Y	29.964	36.331	34.638	56.157	45.016	40,673	35,293	35.949	29.964
C32	**Y**	**N**	**N**	**N**	**N**	**N**	**29.281**	**34.574**	**33.028**	**57.961**	**44.977**	39,775	29,320	**25.612**	**25.612**

Findings

- State-of-the-art best **AI architecture**

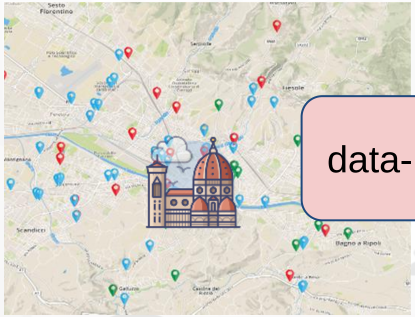
CONV-BI-LSTM



- Relevant **features** in prediction computation

traffic, **datetime**, seasonality, temporal, weather, pollutants

- **validation** performed on real-world road infrastructure



data-missing

MAPE	
0% **miss_rate**	**10%** **miss_rate**
21.34%	23.16%

UNIVERSITÀ DEGLI STUDI FIRENZE

DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DISIT DISTRIBUTED SYSTEMS AND INTERNET TECHNOLOGIES LAB

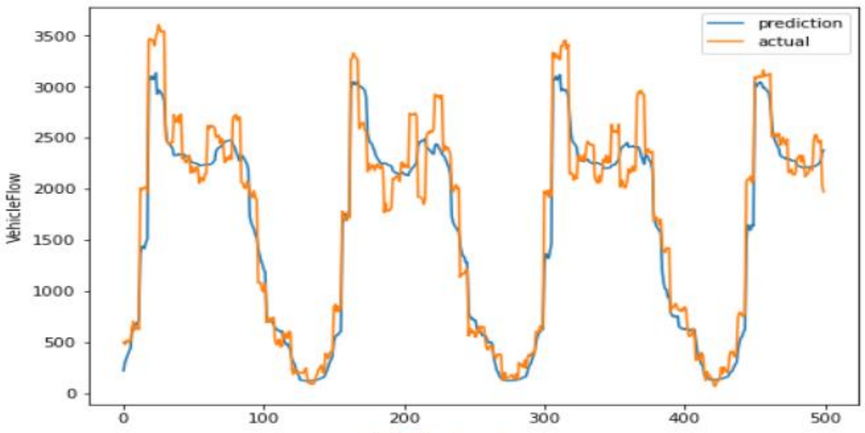
SNAP4CITY

KM4CITY

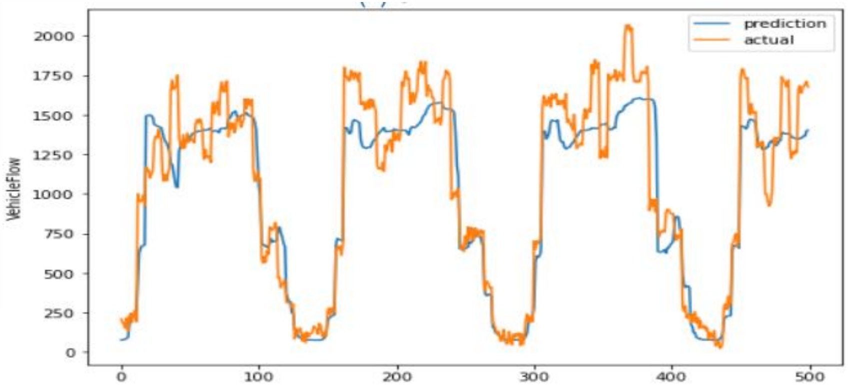
Traffic Flow Prediction via Convolutional Deep Learning

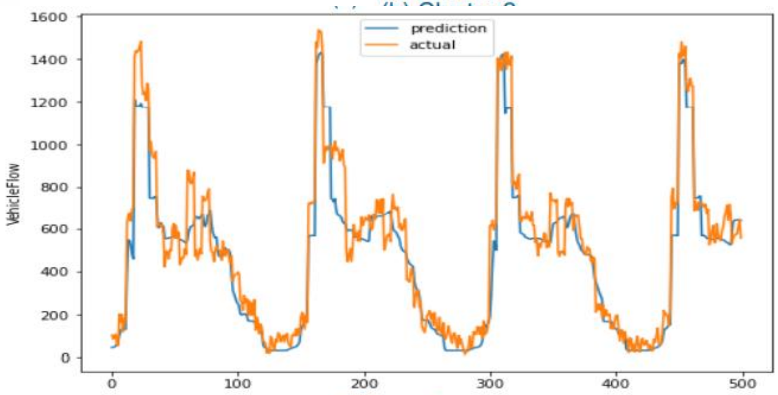
Appendix

Predictions On Representative Sensors



(a) Cluster 1





I Cluster 3

IMPACT OF DATA MISSING ON PRECISION

- **Data missing is an inevitable problem** when dealing with real world IoT sensor networks. Traffic sensors may suffer of problems such as detector malfunction and communication failure.
- The presence of missing data samples in making predictions (execution of the predictive model)may impact on the precision, up to make **impossible to produce the prediction**
- The approach of data imputation can be a valid option to produce surrogate data.
- In this case it has been used an **Hot-Deck** imputation.

IMPACT OF DATA MISSING ON PRECISION

- The robustness has been assessed on the test dataset from 10/02/2020 to 16/02/2020 randomly setting to missing the Vehicle Flow of a percentage of the total dataset based on the missing rates chosen (10%, 25%, 50%, 75%) and then imputing the missing data.
- The imputation strategy proposed to handle missing data reports valid results for the missing rates of 10%, 25%, 50%, 75% on all the representative sensors of the three clusters.

TABLE VIII - DATA MISSING ANALYSIS BASED ON DIFFERENT MISSING RATES ON THE CLUSTERS REPRESENTATIVE SENSORS OF TABLE VII.

Representative sensor	Missing Rate	MAE	MAPE	RMSE	R2
cluster 1 METRO775	0%	161.42	15.35	221.84	0.95
	10%	173.19	16.12	241.86	0.94
	25%	177.36	17.17	258.88	0.93
	50%	176.98	16.77	258.26	0.93
	75%	173.92	16.67	248.51	0.93
cluster 2 METRO707	0%	138.98	23.86	182.48	0.90
	10%	147.49	25.36	194.64	0.88
	25%	146.77	24.90	193.56	0.88
	50%	145.72	24.52	193.34	0.88
	75%	146.10	24.58	193.46	0.88
cluster 3 METRO714	0%	81.86	25.73	117.37	0.89
	10%	83.73	27.99	119.32	0.87
	25%	83.01	27.15	119.11	0.87
	50%	85.00	28.92	122.33	0.87
	75%	82.18	26.89	118.42	0.88

CONV-BI-LSTM Architecture Proposed

The defined **CONV-BI-LSTM** network is made up of 3 components:

- The first component is made up of a **Convolutional 1-dimensional layer** with 48 filters and a kernel size of 16, and a Max Pooling layer of 2x2 and stride equal to 1.
- The second component is the **BI-LSTMs** layers, in particular **6 layers** with **32 units** per layer and **dropout of 0,25**.
- The last one is made of **3 fully connected layers with number of neurons of 32-16-1**. The last one has a sigmoid activation to produce the prediction.

The used optimizer is **Adam Optimizer** with learning rate between **0.005 and 0.008**. **MSE** was selected as the loss function to be monitored during **optimization**. The **batch size** has been set to **512** and the number of epochs was set to a maximum value of 1000, because the training strategy used the **Early Stopping** method with patience parameter set to 100 to determine the optimum epoch number minimizing the MSE of the validation set, restoring the weights of the best model at the end of the learning process.

Features

- One of the goals of this work has been conquer a general understanding above the factors that are more relevant for predicting traffic conditions in the city. Based on the related works, a set of data composed of **temporal variables, traffic-related features**, **weather** information, and **air pollution** has been considered.

Category	Feature	Description
Traffic Trafplus	*Vehicle Flow*	Real number of vehicles recorded every 10 minutes
	AverageSpeed	Average speed of vehicles (Km/h)
	Concentration	Number of vehicles in terms of road occupancy (%)
DateTime	*timeOfTheDay*	Time of the day {1, 144}
	dayOfTheYear	Day of the year {1, 366}
seasonality	*dayOfTheWeek*	Day of the week {1,7}
	Weekend	0 for working days, 1 else
	Year	The year of the observation
Temporal	Previous observation's difference of the previous week ()	the difference between the number of vehicles in the observation day (d) at the time slot t and the number of available bikes during the previous time slot (t-1) of the previous day (d-1)
	Subsequent observation's difference of the previous week ()	the difference between the number of vehicles in the observation day (d) at the time slot t and the number of bikes during the successive time slot (t+1) of the previous day (d-1).
	Previous week observation ()	the number of vehicles of the previous week (d-7) in the same time slot (t).

Features

- One of the goals of this work has been conquer a general understanding above the factors that are more relevant for predicting traffic conditions in the city. Based on the related works, a set of data composed of **temporal variables, traffic-related features**, **weather** information, and **air pollution** has been considered.

Category	Feature	Description
Weather	*Air Temperature*	City temperature one hour earlier than *Time* (°C)
	Humidity	City humidity one hour earlier than *Time* *(%)*
	Pressure	City pressure one hour earlier than *Time* *(millibar mb)*
	Wind Speed	City wind speed one hour earlier than *Time (KM/h)*
AirPoll	*CO*	Concentration of CO one hour earlier than *Time*
	NO2	Concentration of NO2 one hour earlier than *Time*
	O3	Concentration of O3 one hour earlier than *Time*
	PM10	Concentration of PM10 one hour earlier than *Time*
	PM2.5	Concentration of PM2.5 one hour earlier than *Time*
Weather	*Air Temperature*	City temperature one hour earlier than *Time* (°C)
	Humidity	City humidity one hour earlier than *Time* *(%)*

FEATURE CATEGORY IMPORTANCE ANALYSIS

- It has been performed a feature importance analysis using the CONV-BI-LSTM model on the representative sensor of Cluster-1.
- The analysis calculated the MAPEs using all the features except the specific considered category excluding recursively each single feature category
- The DMAPE is defined as the difference of MAPE with respect to the minimum MAPE registered for the CONV-BI-LSTM such as:

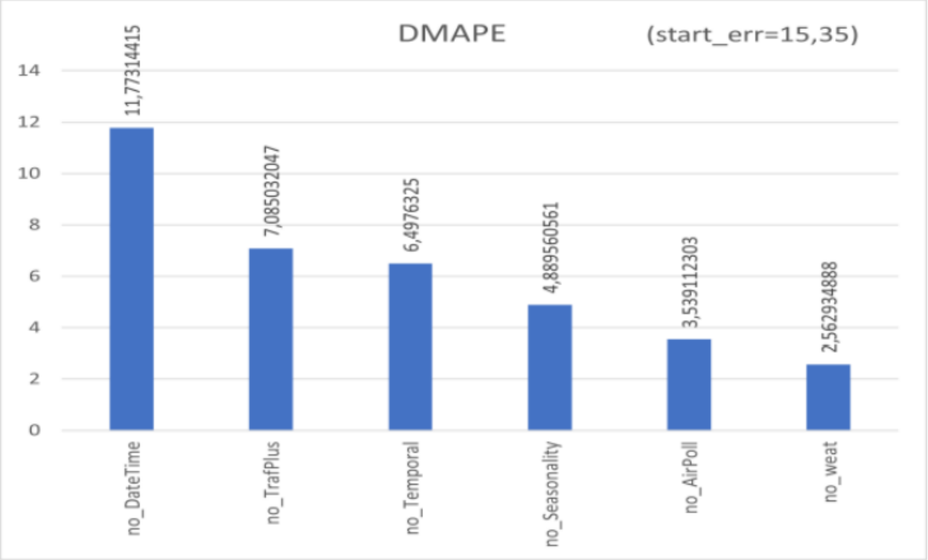
$$DMAPE_i = MAPE_{all-cat_i} - minMAPE$$

Where: i = 1, …, number of categories-1 (all, except the traffic, for a total of 6).

Categories with a higher DMAPE are the most relevant ones, since they do not cause larger differences / errors.

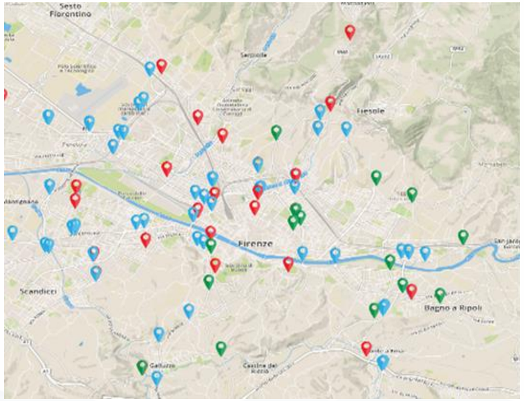
FEATURE CATEGORY IMPORTANCE ANALYSIS

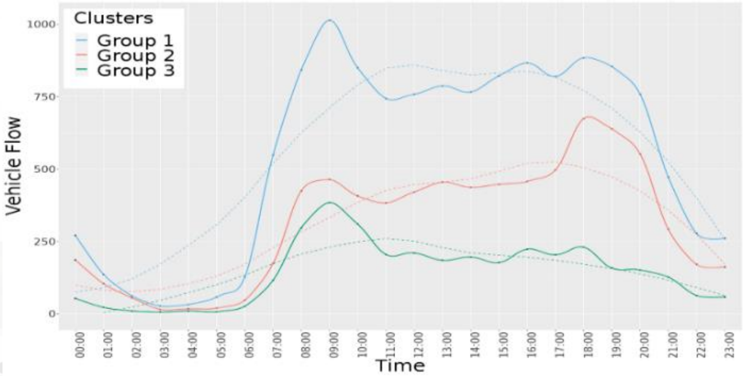
- The feature category with the highest DMAPE is the **DateTime** followed by the **Trafplus**, and the **Temporal feature** category. Additional information on data seasonality for short-term prediction has been ranked 4th , ahead of Air Pollution feature category which in turn beats also Weather features



Clustering

- The clustering has been performed on the basis of the time trend H24, considering the normalized vehicle flow measures.
- The optimal number of clusters turned out to be 3 and it has been identified by using **elbow** criteria
- **K-means** clustering method has been applied to identify clusters
  – The optimal number of clusters resulted to be equal to **3**, and it has been identified by using the **Elbow criteria**





Evaluation Metrics

- **Root Mean Squared Error (RMSE)**

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(obs_i - pred_i)^2}{n}}$$

R-Squared(R2)

- $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{obs_i}$
- $R^2 = 1 - \left(\frac{\sum_{i=1}^{n}(obs_i - pred_i)^2}{\sum_{i=1}^{n}(obs_i - \overline{y})^2}\right)$

Mean Absolute Scaled Error (MASE)

$$q_t = \frac{obs_t - pred_t}{\frac{1}{n-1}\sum_{i=2}^{n}|\,obs_i - \ obs_{i-1}|}$$

$$MASE = mean\,(|q_t|), \qquad t = 1, \dots, n$$

Mean Absolute Error (MAE)

$$MAE = \frac{\sum_{i=1}^{n}|obs_i - pred_i|^2}{n}$$

N	Y	Y	Y	Y	Y	61.579	71.245	77.572	82.634	49.253	60,249	62,308	47.044	47.044
N	Y	Y	Y	Y	N	63.153	71.786	82.539	40.695	47.843	59,814	58,809	35.080	35.080
N	Y	Y	Y	N	Y	61.337	67.098	74.538	48.810	39.500	53,299	54,460	34.383	34.383
N	Y	Y	Y	N	N	59.836	69.497	73.069	51.573	43.162	56,329	51,884	30.699	30.699
N	Y	Y	N	Y	Y	62.579	70.836	78.562	52.011	44.275	57,555	58,888	39.215	39.215
N	Y	Y	N	Y	N	65.235	74.594	71.483	47.473	47.431	61,012	53,134	34.786	34.786
N	Y	Y	N	N	Y	60.648	67.809	65.560	50.268	45.742	56,775	49,816	34.072	34.072
N	Y	Y	N	N	N	65.146	68.154	62.328	52.430	38.560	53,357	46,883	31.439	31.439
N	Y	N	Y	Y	Y	87.662	87.685	96.116	48.239	55.073	71,379	65,153	34.190	34.190
N	Y	N	Y	Y	N	92.600	91.702	97.397	45.643	38.839	65,270	66,528	35.660	35.660
N	Y	N	Y	N	Y	81.587	87.461	91.794	98.094	74.307	80,884	64,471	37.149	37.149
N	Y	N	Y	N	N	88.379	93.141	97.063	48.135	43.457	68,299	66,708	36.354	36.354
N	Y	N	N	Y	Y	89.694	86.278	93.713	47.987	39.295	62,786	64,921	36.130	36.130
N	Y	N	N	Y	N	95.933	90.567	97.018	43.781	55.774	73,170	65,662	34.307	34.307
N	Y	N	N	N	Y	81.423	82.424	84.088	57.282	44.650	63,537	59,109	34.131	34.131
N	Y	N	N	N	N	105.358	88.863	83.781	57.603	39.214	64,038	56,526	29.271	29.271
N	N	Y	Y	Y	Y	66.155	70.646	99.553	71.170	55.400	63,023	73,459	47.366	47.366
N	N	Y	Y	Y	N	70.520	75.833	92.552	64.248	47.862	61,847	63,637	34.723	34.723
N	N	Y	Y	N	Y	68.477	65.479	84.188	74.767	57.531	61,505	60,418	36.648	36.648
N	N	Y	Y	N	N	75.340	68.767	77.498	68.524	49.812	59,289	53,901	30.304	30.304
N	N	Y	N	Y	Y	63.759	69.235	79.907	65.290	54.281	61,758	63,021	46.135	46.135
N	N	Y	N	Y	N	69.726	69.803	71.915	46.227	54.910	62,356	53,797	35.679	35.679
N	N	Y	N	N	Y	64.300	64.550	67.123	56.830	83.575	74,062	49,193	31.264	31.264
N	N	Y	N	N	N	75.239	67.306	69.582	57.716	49.640	58,473	52,173	34.764	34.764
N	N	N	Y	Y	Y	98.969	91.366	112.230	66.553	80.339	85,852	74,096	35.963	35.963
N	N	N	Y	Y	N	101.122	96.404	99.889	60.807	84.501	90,452	66,802	33.715	33.715
N	N	N	Y	N	Y	86.663	88.952	85.747	66.840	94.029	91,490	63,157	40.567	40.567
N	N	N	Y	N	N	86.249	89.043	95.480	59.384	57.615	73,329	63,936	32.392	32.392
N	N	N	N	Y	Y	97.382	91.550	102.559	45.337	43.743	67,646	71,048	39.538	39.538
N	N	N	N	Y	N	98.401	91.546	99.451	60.519	54.513	73,029	70,780	42.109	42.109
N	N	N	N	N	Y	83.403	81.393	84.196	51.608	51.096	66,244	62,604	41.012	41.012
N	N	N	N	N	N	88.430	87.844	85.450	56.906	44.187	66,015	60,793	36.137	36.137