Web Intelligence Network Conference - From Web to Data
Statistic Poland - Arche Dwór Uphagena Hotel
Gdańsk (Poland) - 4 February 2025 - 5 February 2025

Integrating Big Data and Administrative Sources for Estimating Vehicle Mileage and Analyzing Road Traffic Accidents

Marco Broccoli [a] , Silvia Bruzzone [b] and Riccardo Giannini [c]

[a] Istat – Italian National Institute of Statistics - Directorate for Methodology and Statistical Process Design

[b] Istat - Italian National Institute of Statistics - Directorate for Social Statistics and Welfare

[c] Istat - Italian National Institute of Statistics - Directorate for Information Technologies

Presentation Outline

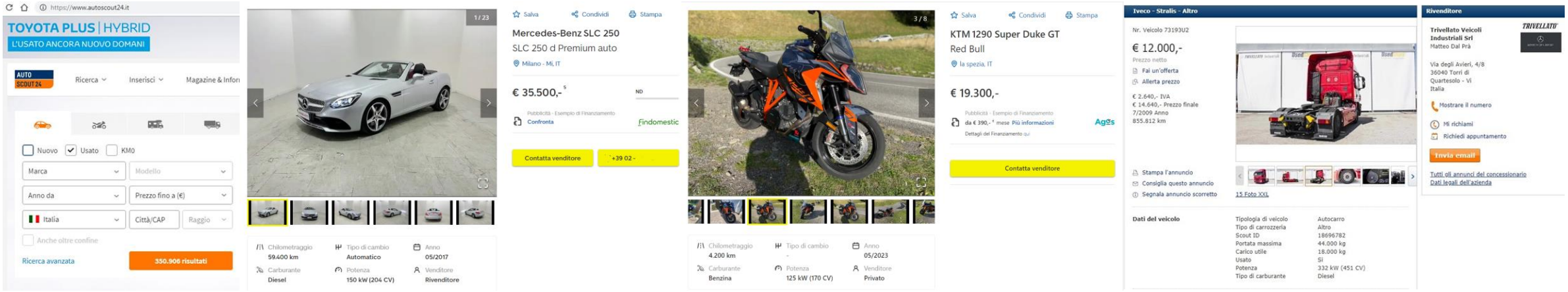
- Project Target
- Identification of the Big Data Source
- Procedural Workflow for Massive Web Scraping
- The Technology Behind an iMacros-Based Macro
- Selection of Vehicle Categories
- Software Architecture of the Project
- Output Generated by Web Scraping
- Methodology Applied for Validation
- Volumes of the Comparative Administrative Data Source
- Verification of Results
- Conclusions

Project Target

- ✓ The goal is to estimate the average mileage covered by vehicles listed for sale, segmented by type, emission class, fuel type, province (or city of sale), and other statistically relevant attributes.
- ✓ This data will be compared with the variables present in the Public Motor Vehicle Registry (PRA) and the Vehicle Inspection Archive, provided by the Ministry of Infrastructure and Transport (MIT).
- ✓ Estimating vehicle kilometers traveled (VKT) on the national road network is part of a broader project. The ultimate aim is to estimate traffic flows and the real exposure risk rates for road accidents.
- ✓ The project also seeks to compare and integrate data from administrative sources and Big Data to test the potential and validity of both sources. The added value derived from merging these datasets will be utilized.
- ✓ The administrative sources are already documented in Istat's QRCA (Quality Report Card for Administrative data) system, accessible with specific data processing authorizations.

Identification of the Big Data Source

- ✓ The proposed approach is optimal since the Autoscout24 and TruckScout24 databases include a sample of vehicles, motorcycles, and heavy vehicles for sale, even within the first four years of registration. These vehicles are not yet subject to mandatory inspections. For heavy vehicles, the timeframe is limited to the first year.

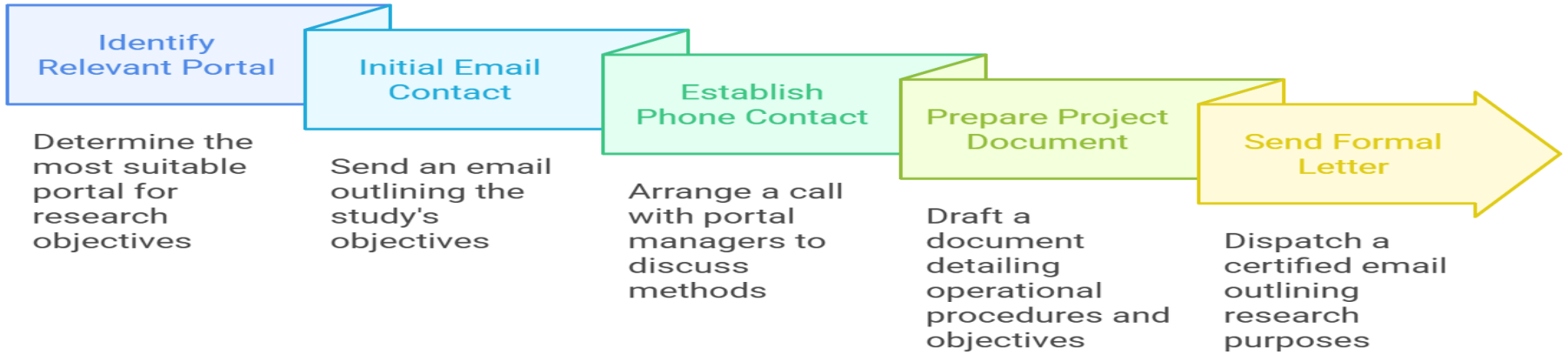


- ✓ In the license plate matching process between the two datasets to estimate average monthly mileage, new vehicles not yet inspected and mopeds are excluded.
- ✓ The Big Data source complements the missing information by estimating mileage for such vehicles.

Procedural Workflow for Massive Web Scraping

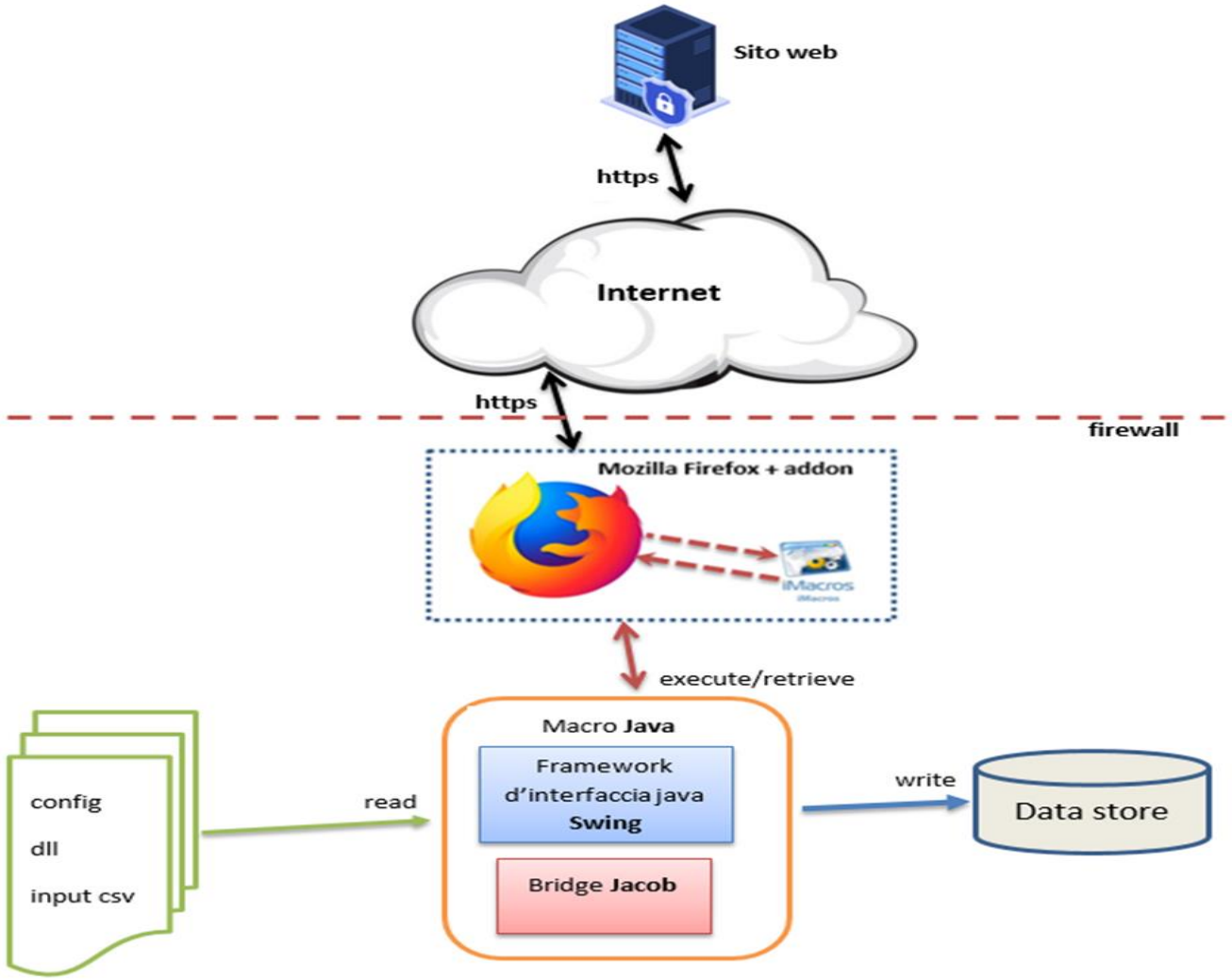
The steps for defining the procedural workflow include:

- ✓ Identifying the most relevant portal for the research objectives.
- ✓ Initial contact via email, outlining the study's objectives.
- ✓ Upon receiving a response, establishing a phone contact with the portal's marketing and IT managers, followed by meetings to discuss operational methods.
- ✓ Preparing a project document detailing the operational procedures, scheduling web scraping on the portal, and specifying the statistical objectives. The document includes the IP address from which queries will be executed. Adding Istat's IP to the portal's whitelist is essential to prevent access blocks due to perceived scraping threats.
- ✓ Sending a formal letter via certified email (PEC) outlining the research purposes.



Technology Behind an iMacros-Based Macro

- ✓ Each macro is developed in Java using the proprietary iMacros software (enterprise version).
- ✓ Java version 8 is employed.
- ✓ The macro uses either the proprietary iMacros browser or Mozilla Firefox (with the iMacros add-on) to access websites via HTTPS.
- ✓ If needed, the macro can interact with users via the Swing framework.
- ✓ The Jacob bridge enables interaction between the Java macro, the browser, and the add-on.



Selection of Vehicle Categories

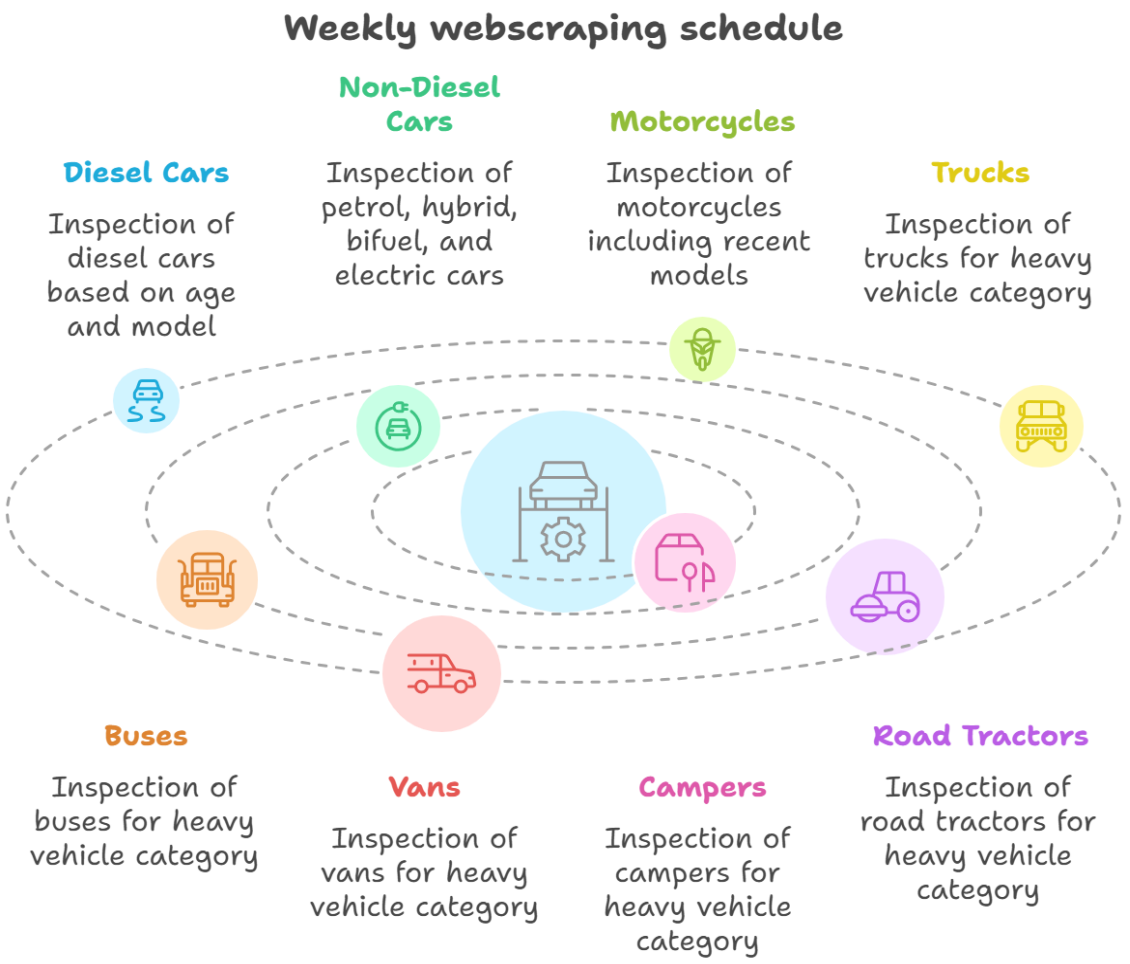
Twelve macros are executed weekly:

For **Autoscout24.it** (*light vehicles*):

- Diesel cars inspected
- Recent diesel cars
- Non-diesel cars (petrol, hybrid, bifuel, and electric) inspected
- Recent non-diesel cars
- Motorcycles inspected
- Recent motorcycles

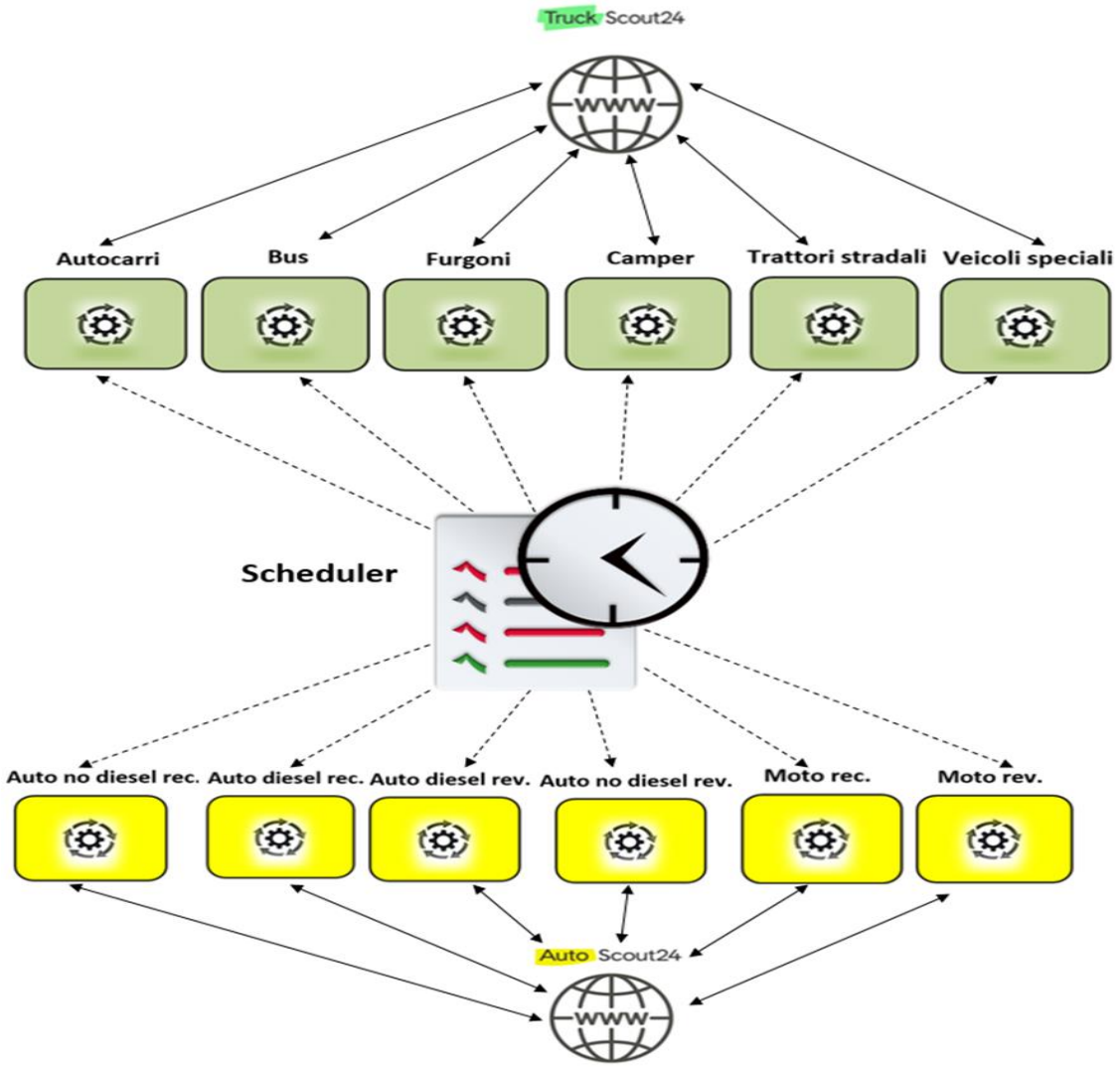
For **Truckscout24.it** (*heavy vehicles*):

- Trucks
- Buses
- Vans
- Campers
- Road tractors (motor units)
- Special vehicles



Software Architecture of the Project

- ✓ Data collection is fully automated.
- ✓ The system scheduler executes all macros based on predefined timings.
- ✓ Two macro batches (six instances each) handle web scraping for Autoscout24 and TruckScout24.
- ✓ Extracted sale ads:
  - 319,895 for heavy vehicles
  - 778,931 for light vehicles



Software Architecture of the Project

- ✓ Each macro generates a CSV file containing data from the selected ad pages.
- ✓ Each ad and its details are summarized in a row of the matrix shown as an example.
- ✓ All CSV files are stored in a relational database for statistical analysis.

CHILOMETRI	IMMATR.	POTENZA	SEZIONE	TIPO	SCOUT_ID	ALIM.	EMIS.	CITTA_VENDITORE	DATA_SCRITTURA	NUM_PAG.
298.500	5/1999	294 kW (400 CV)	Autocarro	Betoniera	18781907	Diesel	Euro2	25017 Lonato del Garda - B	27/03/2019 16.56	5
314.400	9/2008	332 kW (451 CV)	Autocarro	Autocarro con cassone ribaltab	18781836	Diesel	Euro4	25017 Lonato del Garda - B	27/03/2019 16.57	5
115.300	1/2013	135 kW (184 CV)	Autocarro	Pianale telonato	18781798	Diesel	Euro5	25017 Lonato del Garda - B	27/03/2019 16.57	5
422.000	9/2001	125 kW (170 CV)	Autocarro	Pianale telonato	18781678	Diesel	Euro2	25017 Lonato del Garda - B	27/03/2019 16.58	5
270.000	6/2003	290 kW (394 CV)	Autocarro	Betoniera	18781622	Diesel	Euro3	25017 Lonato del Garda - B	27/03/2019 17.00	5
1.260.200	7/2004	335 kW (455 CV)	Autocarro	Pianale telonato	18781612	Diesel	Euro3	25017 Lonato del Garda - B	27/03/2019 17.00	5
667.000	10/2005	335 kW (455 CV)	Autocarro	Autocarro con cassone ribaltab	18781600	Diesel	Euro3	25017 Lonato del Garda - B	27/03/2019 17.01	5
366.700	4/2009	353 kW (480 CV)	Autocarro	Autocarro con cassone ribaltab	18781578	Diesel	Euro5	25017 Lonato del Garda - B	27/03/2019 17.02	5
400.000	4/2003 ,	335 kW (455 CV)	Autocarro	Autocarro con cassone ribaltab	18781287	Diesel	Euro3	29010 Pontenure - Piacenza	27/03/2019 17.03	5
222.635	12/2001		Autocarro	Pompa per calcestruzzo	18780589	Diesel	Euro3	25030 Lograto - Brescia	27/03/2019 17.05	5
554.916	2/2012	353 kW (480 CV)	Autocarro	Telaio intercambiabile	18782183	Diesel	Euro5	00155 Roma - Rm	27/03/2019 17.06	5
185.523	10/2010		Autocarro	Altro	18780463	Diesel	Euro5	31040 Pederobba - Tv	27/03/2019 17.07	5
585.000	8/1999	228 kW (310 CV)	Autocarro	Autocarro con cassone ribaltab	18779244	Diesel	Euro2	25017 Lonato del Garda - B	27/03/2019 17.08	5
280.000	7/2003	110 kW (150 CV)	Autocarro	Autoc. scarrabile	18779192	Diesel	Euro3	25017 Lonato del Garda - B	27/03/2019 17.09	5
109.900	11/2002	324 kW (441 CV)	Autocarro	Pompa per calcestruzzo	18779177	Diesel	Euro3	25017 Lonato del Garda - B	27/03/2019 17.10	5
816.000	8/2006	355 kW (483 CV)	Autocarro	Furgone silos	18779140	Diesel	Euro3	25017 Lonato del Garda - B	27/03/2019 17.10	5
495.900	6/2015		Autocarro	Autotelaio	18778800	Diesel	Euro6	70026 Modugno - Bari	27/03/2019 17.11	5
168.500	6/2015		Autocarro	Autotelaio	18778772	Diesel	Euro6	70026 Modugno - Bari	27/03/2019 17.12	5
321.412	1/2013	140 kW (190 CV)	Autocarro	Furg. trasp. refrigerato ISO	18779678	Diesel	Euro5	10024 Moncalieri - To	27/03/2019 17.15	6
765.215	3/2003	158 kW (215 CV)	Autocarro	Furg. trasp. refrigerato ISO	18779680	Diesel	Euro3	10024 Moncalieri - To	27/03/2019 17.17	6
374.533	6/2013	183 kW (249 CV)	Autocarro	Autotelaio	18779681	Diesel	Euro5	10024 Moncalieri - To	27/03/2019 17.17	6
435.793	6/2013	183 kW (249 CV)	Autocarro	Autotelaio	18779682	Diesel	Euro5	10024 Moncalieri - To	27/03/2019 17.18	6
390.000	6/2003		Autocarro	Autoc. scarrabile	18776177	Diesel	Euro3	29010 Pontenure - Piacenza	27/03/2019 17.26	6
297.000	8/2012	183 kW (249 CV)	Autocarro	Altro	18775588	Diesel	Euro4	10024 Moncalieri - To	27/03/2019 17.27	6
464.000	6/2007		Autocarro	Furg. trasp. refrigerato ISO	17924955	Diesel	Euro4	20090 Cusago - Mi	27/03/2019 17.27	6
450.000	2/2007		Autocarro	Pianale telonato	18695162	Diesel	Euro4	80024 Cardito - Na	27/03/2019 17.28	6
326.432	6/2005		Autocarro	Altro	18775686	Diesel	Euro2	36016 Thiene - Vi	27/03/2019 17.29	6

Methodology Applied for Validation

- ✓ Since 2020, the inclusion of vehicle license plate data in the Vehicle Fleet Archive has enabled the linkage of an association key with data from the vehicle inspection source. This includes the total kilometers traveled and the corresponding reference dates (registration and inspection) for each period between 2014 and 2021.

- ✓ Current regulations for vehicle inspections in Italy is biennial for light vehicles and after the first four years for the initial inspection.

- ✓ For heavy vehicles, annual inspections provide a more robust validation of the data consistency through the integration of the two administrative databases, offering comprehensive coverage of the phenomenon.

- ✓ The comparison between the two methods is based on exhaustive data for vehicles that have undergone at least one inspection, meaning those registered before 2017.

- ✓ For vehicles that have not yet undergone inspections, the integration of Big Data sources offers a reliable estimate, compensating for the absence of information on total kilometers traveled.

The volumes of the comparative administrative source

Vehicle Category	Less than 2 Years	Between 2 and 5 Years	Between 5 and 10 Years	Between 10 and 20 Years	Over 20 Years	Total
Passenger Transport Vehicles						
AB - Buses	6.936	15.377	14.863	40.177	22.530	99.883
AC - Motorhomes	1.311	20.438	28.403	126.574	109.864	298.389
AP - Mixed-Use Vehicles	1	14	52	7.328	439.690	447.085
AV - Passenger Cars	3.526.322	7.451.583	7.263.123	14.765.581	6.264.180	39.270.789
Goods Transport Vehicles						
AM - Trucks	317.884	662.193	564.894	1.531.516	1.145.231	4.221.718
AS - Special Vehicles	33.477	72.957	69.635	162.308	127.971	466.348
TS - Road Tractors	21.744	53.184	33.665	61.543	25.333	195.469
Motor Vehicles						
MC - Motorcycles	444.684	761.540	937.214	2.809.591	2.031.293	6.984.322
MM - Motor Tricycles	638	3.184	6.306	29.690	176.953	216.771
MP - Mixed-Use Motor Vehicles			2	16	66	84
MZ - Motorcycles with Sidecar	3.810	4.525	1.965	638	8.274	19.212
QC - Quadricycles	2.943	10.805	21.115	70.613	9.973	115.449
Trailers and Unclassified Vehicles						
RM - Trailers	24172	61696	40487	138750	149693	414798
Unclassified			1	4	18	23
Total	4395721	9117496	8981725	19744329	10.511.069	52.750.340

Year Vehicle Inspection	2014	2015	2016	2017	2018	2019	2020	Total
Four Wheels Vehicles	4.200.525	10.985.000	10.890.578	13.580.702	14.443.370	14.877.732	13.795.060	82.772.967
Motorcycles	187.219	1.013.243	1.095.757	1.417.395	1.480.989	1.522.451	1.307.941	8.024.995
Total	4.387.744	11.998.243	11.986.335	14.998.097	15.924.359	16.400.183	15.103.001	90.797.962

Verification of Results

- ✓ Analysis of the annual average mileage for light vehicles shows similar distributions across both sources.

- ✓ Notably, administrative data methods yield lower mileage averages for older vehicles due to biennial inspections, compared to annual averages between registration and sale dates.

Average Km Traveled by Year of Registration							
Passenger Cars	Recent	First Inspection	Between 5 and 10 Years	Between 10 and 20 Years	Over 20 Years	Average	Total Vehicles
Administrative Source		15.728	13.515	11.328	6.802	12.724	34.485.667
Big Data Source	15.662	17.110	15.646	11.689	4.925	14.870	476.045
Diesel Vehicles	Recent	First Inspection	Between 5 and 10 Years	Between 10 and 20 Years	Over 20 Years	Average	Total Vehicles
Administrative Source		18.798	15.846	13.804	11.972	15.938	15.913.065
Big Data Source	16.389	20.191	17.589	14.195	8.550	17.867	116.196
Motorcycles	Recent	First Inspection	Between 5 and 10 Years	Between 10 and 20 Years	Over 20 Years	Average	Total Vehicles
Administrative Source		4.118	3.349	2.712	1.911	3.064	4.653.892
Big Data Source	5.371	4.307	3.768	2.910	1.584	3.601	226.458

Conclusions

- ✓ The project aims to develop models to estimate the kilometers traveled by vehicles across the national territory for the purpose of defining the indicators of “average daily theoretical vehicles” (V.T.M.G.), which represent the number of vehicles traveling the entire road network daily, calculated as the ratio of kilometers traveled on a road segment to its length in kilometers, multiplied by the number of days. This indicator measures the usage level of the road network, highway, or specific segment. At the end of the process, it will be possible to provide a more accurate estimate of road users' risk exposure for different vehicle categories.
- ✓ Together with road accident indicators based on kilometers traveled—a novel element compared to traditional metrics relying on resident population or vehicle fleets—this approach eliminates part of the effects of the mobility component of the phenomenon and allows for a more adequate territorial comparison. Coupled with other projects aimed at localized mapping of traffic flows on specific areas, road types, or individual segments, this constitutes the goal of precisely identifying risk exposure for different user categories.
- ✓ Additionally, by profiling vehicles involved in road accidents—an aspect never previously considered in road accident analysis—and using matching with vehicle license plates, it will be possible to categorize the involved vehicles based on their average distances traveled.

Web Intelligence Network Conference - From Web to Data
Statistic Poland - Arche Dwór Uphagena Hotel
Gdańsk (Poland) - 4 February 2025 - 5 February 2025

Integrating Big Data and Administrative Sources for Estimating Vehicle Mileage and Analyzing Road Traffic Accidents

Marco Broccoli [a] - broccoli@istat.it – Project manager
Silvia Bruzzone [b] - bruzzone@istat.it – Thematic Content Expert
Riccardo Giannini [c] - rigianni@istat.it – IT and Web Scraping Expert

Istat | Istituto Nazionale di Statistica