BIG DATA AND ROAD SAFETY: OPEN STREETMAP AND THE TERRITORIAL ANALYSIS OF ACCIDENTS[1]

Marco Broccoli, Silvia Bruzzone

Abstract. Road safety is one of the main challenges for sustainable mobility policies, aiming to reduce the number of accidents and their consequences. Analysing road accidents is essential to identify risk factors and to develop prevention strategies, based on data. The aim of this study is threefold. First, to utilize OpenStreetMap (OSM) data to calculate road accident, mortality, and injury indices by correlating them with the length of road lanes (in meters). Second, to conduct a territorial analysis to identify high-risk areas, thereby supporting road safety planning and third, to enhance national statistical information by estimating accident involvement probabilities, with the ultimate goal of determining real traffic flows (vehicles/km) and actual risk exposure rates. The approach uses an integrating geographic and statistical data using GIS techniques. The researchers implement a spatial join algorithm to overlay information layers derived from OSM and traffic points (PoT).
The analysis includes a new classification of road segments, updated to 2021, and the application of the "Ranker" tool to generate synthetic indicators. Accident data, provided by Istat and other administrative sources, are georeferenced and analysed to highlight territorial variations in risk distribution.
The main innovation of this study lies in the use of Big Data from OSM for statistical purposes, aligning with Trusted Smart Statistics (TSS) initiatives. The integration of geographic and statistical data overcomes the limitations of traditional risk measures based on resident population or vehicle ownership. Furthermore, the introduction of traffic points refines risk indicators, providing a more detailed framework for accident prevention on a territorial scale.

1. Introduction

Road safety remains a paramount challenge within sustainable mobility policies, with concerted efforts aimed at reducing the number of accidents and their severe consequences. Road Safety Performance Indicators (RSPI), as recommended by European Commission programs for EU countries, offer a multidimensional approach to investigating accidents, considering roads, vehicles, and individuals involved. Preventing road trauma on public roads is a core responsibility for governments, their agencies, and stakeholders, necessitating a common and shared commitment. The scale

[1] Marco Broccoli edited paragraphs 2, 2.1, 2.2, 2.3, 2.4, 2.5, 3, 3.1, 3.2, 3.3, 3.4 and 3.6; Silvia Bruzzone edited paragraphs 1, 3.5 and 4.

of the road safety challenge and the diversity of its impacts underscore the importance of exploring synergies among decision-makers within the road network.

In Italy, road accidents resulting in death or injury continue to pose a significant public health and social burden. Data for the years 2010-2023 illustrate ongoing trends, with a target for 2030 aiming at halving the number of deaths[2]. While there has been a general downward trend in accidents and injuries over the past decade, the number of fatalities remains a critical concern, with an estimated 3,030 deaths in 2024. This perspective highlights the urgent need for effective, data-driven prevention strategies.

Traditionally, analysts calculate road fatality and accident rates based on denominators such as the resident population or the vehicle fleet in each province of registration. However, a clear information bias exists regarding the appropriate reference denominators for constructing robust statistical indicators linked to road accidents (Broccoli and Bruzzone, 2021). Resident population, often used as a common proxy for the population exposed to risk in a specific geographical area, is not always an appropriate solution. This is especially true given the seasonal nature of road accidents and their concentration in specific locations or periods, which means resident population figures, do not accurately reflect the actual population present at the time and place of an event. Similarly, while the vehicle fleet provides more targeted information, it still suffers from a deductible bias due to the mobility of road users, failing to account for vehicles transiting through an area.

These traditional indicators, though more easily accessible, are therefore inherently limited. The high mobility of individuals for work, leisure, family needs, and commuting creates a substantial distortion when attempting to assess accident risk in specific geographical areas. In contrast, this study leverages the length of the road network (carriageway length in meters) from Open Street Map (OSM) as a more stable and geographically anchored denominator. The road infrastructure, unlike population or vehicle fleet, remains a constant physical presence within a territory, offering a more consistent basis for risk assessment. The ultimate goal is to enhance national statistical information by estimating the probability of accident involvement, accounting for different risk exposures, and eventually estimating effective traffic flows (vehicles/km) on the national road network.

This research aligns with Istat's initiatives in Experimental Statistics and the pursuit of "Trusted Smart Statistics" (TSS). Since 2019, Istat has been developing experimental statistics that include road accident, mortality, and harmfulness rates, comparing traditional measures with those based on roadway segment length from OSM. The final aim for these experimental statistics is to publish the documents as official statistics, with a continuous and planned timetable by 2026. Beyond academic advancement, a

[2] European Commission. EU Road Safety Policy Framework 2021-2030 - Next steps towards “Vision Zero”. Brussels 19.6.2019, SWD (2019) 283 final. Link: https://transport.ec.europa.eu/system/files/2021-10/SWD2190283.pdf

primary driver of this research is the development of robust synthetic indicators that can serve as actionable tools for policymakers. Such indicators are crucial for enabling more informed decision-making in the programming of preventive actions, the strategic modernization of road infrastructures, and, where appropriate, the targeted enforcement of driving behaviours to enhance overall road safety. This paper presents the methodology and main results of using OSM data for a territorial analysis of road accidents, highlighting the advantages of innovative denominators and synthetic indicators.

2. Data and Methods

The approach integrates geographic and statistical data using Geographic Information Systems (GIS) techniques and OSM as a key Big Data source.

2.1 Open Street Map Data

OSM is a collaborative project aimed at creating free, editable content maps of the world. It provides a vast collection of geographical data, including detailed road networks, with the primary purpose of creating maps and cartography. A key feature of OSM data is its free license (Open Database License), allowing for its use for any purpose with the only constraint of mentioning the source. A global community contributes data using GPS devices, aerial photography, and other free sources. Most Android and iOS GPS navigation software (e.g., WisePilot, Maps.me, NavFree) are powered by OSM.

For this study, key OSM vector layers, which are daily updated and freely downloadable, include:

- Road graph (detailing road segments and their characteristics);
- Traffic Points (PoT), indicating locations on road segments where traffic intensity.

2.2 GIS Techniques and Data Integration

A spatial join algorithm integrates the data by linking Istat's census geography with the OSM road network. This process overlay the two vector layers, enriching Istat administrative features (e.g., localities) with OSM road segment attributes based on spatial location. The procedure generates a unified dataset that incorporates both administrative boundaries and detailed infrastructure characteristics through the key-reference-by-position algorithm.

2.3 Classification of Road Segments and Localities

To build road accident indicators with denominators represented by road segment length from OSM, Istat uses a "bridge coding table". This table systematically classifies OSM road segments according to Istat survey on road accidents classification and Istat Population Census Localities.

The classification considers:

- OSM road segment types (e.g., motorway, trunk, primary, secondary, tertiary, residential, service) (Table 1).
- Istat Population Census Localities types: Urban areas, Small inhabited areas, Productive areas, widespread houses;
- Road accidents survey road types (e.g., Motorway, Urban Road and Rural Road).

Table 1 – *OSM road segments classification (a).*

Road type	Road type description
Secondary link	The link roads (slip roads/ramps) leading to/from or from/to a secondary road or lower class highway.
Tertiary	Roads of local rank. They connect smaller municipalities together. In urban areas, they are side roads to primary and secondary roads with a medium flow of traffic.
Tertiary link	The link roads (slip roads/ramps) leading to/from or from/to a tertiary road or lower class highway.
Unclassified	Classification for some extra-urban road.
Residential	Roads in a residential area, which serve as an access to housing, without function of connecting settlements.
Living Street	Residential road where pedestrians have legal priority over cars, speeds are very low.
Pedestrian	Pedestrian areas (roads or squares in urban areas), accessible mainly or to pedestrians.
Service	Access roads or internal service areas, beaches, camping, industrial areas, shopping centers, parking places etc.
Track	Roads for mostly agricultural or forestry uses
Cycle way	Cycle paths on dedicated carriageway, mainly or exclusively for cycling tourism.
Footway	Paths mainly/exclusively for pedestrians. This includes walking urban tracks, paths in a public park and footpaths
Path	Paths not structured for a public use
Steps	Stairs in steps, exclusively accessible by pedestrians
Unknown	Not classified

(a) Road Segments by Open Street Map – update 1/1/2024.

The study adopts a new analytical classification with respect to the first release (Broccoli and Bruzzone, 2021), using a more refined technique for attributing individual road segments, approximately 3.5 million in total from OSM to the Istat classification

groups (Table 2). The operational criterion applied provides the roads classification, through the textual analysis of the Name and Reference attributes, according to the different classes of road segment and spatial attribution of the location type.

Table 2 – *Bridge coding table between roads segments classification by OSM, localities and road type (a).*

Road Segments classification by OSM	Localities at Census 2011			
	Urban areas + Small		Productive areas + Wide	
	Road Localisation by Road accidents survey			
	Motorways	Urban Roads	Motorway	Rural Roads
Motorway	x		x	
Trunk	x		x	
Primary		x		x
Secondary		x		x
Tertiary		x		x
Unclassified		x		x
Residential		x		x
Living Street		x		x
Motorway Link	x		x	
Trunk Link	x		x	
Primary Link		x		x
Secondary Link		x		x
Tertiary Link		x		x
Service		x		x
Unknown		x		x

(a) *Istat computing*

2.4 Road Accident Indicators and Traffic Point (PoT) Weighting

The study calculates road accident, mortality, and injury indices. The innovation lies in correlating these with the length of road lanes (in meters) by driving direction from OSM. As a further refinement, the study used additional information on Traffic Points (PoT) on road segments, downloaded from the OSM detection system. Since the 2021 edition of experimental statistics, Istat researchers propose road accident indicators "weighted" with information on traffic intensity. This information considers the kilometres of roadway with the presence of a traffic point as a discriminating element, aiming to better approximate actual traffic exposure. The ultimate objective is to estimate real traffic flows (vehicles/km).

2.5 Synthetic Indicators: The Ranker Tool and MZ Method

For the analysis and comparison of the synthetic indicators, Istat uses two tools developed in-house:

- Ranker Tool desktop software: (http://www.istat.it/en/tools/methods-and-it-tools/analysis-tools/ranker)
- i.Ranker web application: (https://i.ranker.istat.it)

Three synthesis methods are evaluated: the arithmetic mean of z-scores (MZ), the Relative Index method (MR), and the Mazziotta-Pareto Index method (MPI; De Muro *et al.*, 2010). For the analysis and comparison of the synthetic indicators, Istat uses tools developed in-house, specifically the "Ranker" and "COMIC" (COMposite Indices Creator) software. Three synthesis methods are evaluated: the arithmetic mean of z-scores (MZ), the Relative Index method (MR), and the Mazziotta-Pareto Index method (MPI; De Muro *et al.*, 2010).

The robustness analysis entailed a comparative assessment of these methods using the COMIC software to verify their internal consistency. Specifically, the stability of the results was measured by analyzing the mean and standard deviation of the rank "shifts" produced by each method. The results demonstrated that both MZ and MPI offered the highest robustness (i.e., minimal rank volatility). Ultimately, the MZ method was selected as the primary criterion due to its computational transparency and ease of interpretation for non-statistical stakeholders.

Regarding the construction of the indicator, the MZ method operates by first standardizing the elementary indicators (accident, mortality, and injury rates) to z-scores (with zero mean and unit variance) to ensure comparability across different scales. Subsequently, the arithmetic mean of these standardized values is calculated to generate the final synthetic index, allowing for consistent territorial comparisons.

3. Main results

3.1. Comparing Rates: Road Length vs Population Denominators

Focusing on 2023 data, road accident indicators by road length (number of accidents, vehicles involved, deaths, and injuries per 1000 km of carriageway in the province) reveal the maximum exposure to risk for motorways and urban roads, primarily stated in main cities. Comparing the road mortality rates with those calculated using resident population as a denominator shows that the ranking of provinces is completely different. For example, Milan and Rome, which rank high for road mortality risk (1st and 4th, respectively) based on road length, drop to 88th and 60th positions out of the resident population. For motorways specifically, their positions shift from 3rd and 11th (by road length) to 17th and 33rd (by population). This clearly demonstrates the distortion introduced by population-based denominators, which do not account for the actual use of the road network. The same results arise even more clearly analyzing the synthetic index provinces ranking (Figure 1, 2).

Figure 1 -*Synthetic Indices: MZ Rank by road resident segments length*

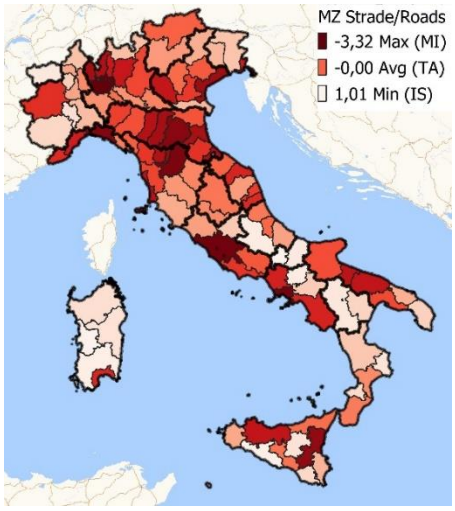
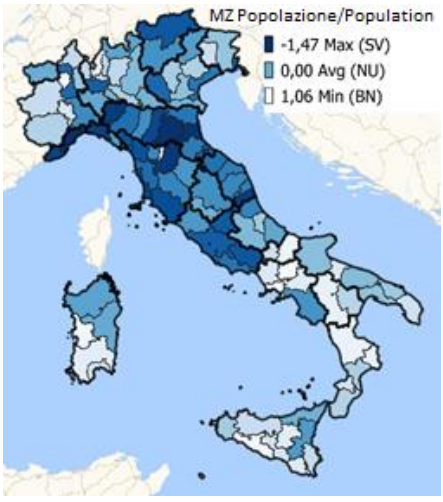


Figure 2 - *Synthetic Indices: MZ Rank by population*



Source: Istat computing on Istat Road Accidents survey (2023), OSM data (1/1/2024) and Istat resident population (1/1/2023).

3.2. Detailed Provincial Rankings and the Impact of Denominators

The analysis of provinces exhibiting the highest and lowest road safety starkly reveals the profound impact of the chosen denominator. The list of the five "worst-performing" provinces, for instance, undergoes significant alterations when switching from road length-based indicators to those reliant on resident population or vehicle fleet. This volatility underscores the critical need to select denominators that accurately reflect risk exposure, as traditional metrics can lead to divergent and potentially misleading policy priorities (Table 3).

Table 3 - *Best and Worst Z Road graph performance. Year 2023 (a).*

	Best Z Road graph performance. Year 2023					
	Ranking indicators by road length		Ranking indicators by vehicles fleet		Ranking indicators by resident population	
Best	Benevento	0.883	Biella	0.780	Biella	0.633
	Sud Sardegna	0.919	Agrigento	1.030	Oristano	0.745
	Agrigento	0.929	Trento	1.045	Napoli	0.767
	Oristano	0.999	Benevento	1.046	Agrigento	1.011
	Isernia	1.017	Aosta	1.330	Benevento	1.062

Table 3 (cont.) - *Best and Worst Z Road graph performance. Year 2023 (a).*

Worst Z Road graph performance. Year 2023						
	Ranking indicators by road length		Ranking indicators by vehicles fleet		Ranking indicators by resident population	
Worst	Milano	-3.330	Bologna	-1.602	Savona	-1.475
	Monza e Brianza	-2.536	Genova	-1.114	Bologna	-1.252
	Roma	-2.085	Savona	-1.108	Ravenna	-1.088
	Napoli	-1.768	Piacenza	-1.072	Firenze	-1.067
	Firenze	-1.720	Ravenna	-1.026	Piacenza	-0.991

(a) A Z-score measures the distance between a data point and the mean using standard deviations. Z-scores can be positive or negative. The sign tells you whether the observation is above or below the mean. The research use the negative polarity to represent the higher risk.

3.3. Correlation Analysis: Highlighting Methodological Differences

The correlation matrix results illustrate the distinctiveness and innovative strength of using road length as a denominator (Table 4).

Table 4 - *Correlation Matrix between indicators (analytical classification of road segments).*

Ranks	Road segment (Analytical classification)	Population	Vehicle Fleet
Road segment	1.0000	0.3794	0.4830
Population	0.3794	1.0000	0.8843
Vehicle Fleet	0.4830	0.8843	1.0000

The low correlation coefficients between indicators based on road segment length and those based on resident population (0.3794) or vehicle fleet (0.4830) empirically confirm the significant differences in risk assessment these approaches yield. This divergence underscores how denominators tied to human mobility (population, vehicle fleet) – influenced by work-related travel, leisure trips, family obligations, and daily commuting – introduce a considerable bias in risk perception. The road network's physical extension, on the other hand, provides a more objective and territorially consistent reference for evaluating accident phenomena. The application of different weighting criteria leads, in fact, to very divergent results. The road accident indicators referred to road length by province, therefore, seem to lead to a better result for the risk measure of the road accidents and are closer to the values by traffic flows data.

3.4. Synthetic Indicators (MZ Method) and Cartographic Representation

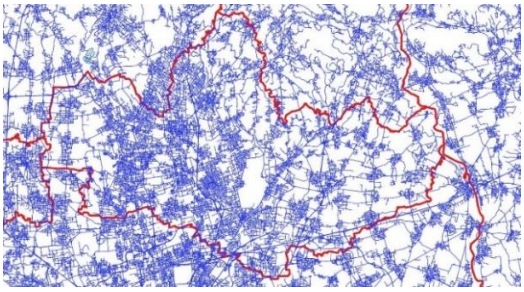
Synthetic indicators using the MZ method provide a composite view of road safety risk. Cartographic representations, applying a quantile method in color classes (higher values highlighted with a more intense tone), visualize these territorial differences

These maps clearly show different risk patterns depending on the normalization basis, reinforcing the utility of road-length based synthetic indicators for policymakers. This refined understanding moves beyond simple incident counts, offering a strategic basis for prioritizing infrastructure upgrades, tailoring public awareness campaigns, and guiding law enforcement.

3.5. Case Studies: Illustrating Denominator Impact

The province of Savona serves as a compelling illustration of the distortions caused by traditional denominators. A significant network of connecting infrastructures and many tourist destinations characterize Savona, leading to substantial traffic volumes from transit and tourism. Consequently, its resident population is relatively small when compared to the actual traffic volumes and the extent of its road network. When road accident risk in Savona is assessed using resident population as a denominator, a potential overestimating of the local risk effect can affect the resulting indicators, because the denominator does not capture the large transient, non-resident road user population. Indicators based on road network length offer a more stable and geographically pertinent assessment of risk in such territories (Figure 3).

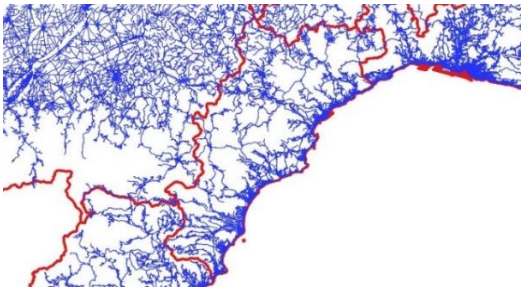
Figure 3 - *Monza e della Brianza road layout by OSM graph. Year 2023.*



Conversely, the province of Monza and Brianza exemplifies a different scenario. An extremely high urban concentration and pervasive urbanization characterize this territory, resulting in a dense road network, intensively utilized through all geographical area. In such a context, the indicator normalized by road carriageway length reaches maximum risk values. This suggests that the sheer density of traffic and interactions on a highly utilized, even if geographically limited, road network creates a heightened risk environment. While population-based indicators might also show high risk, the road length metric emphasizes the intensity of risk concentrated on the existing infrastructure. This is particularly sensitive for identifying areas where infrastructure is under significant pressure, a critical factor in densely urbanized settings. The contrasting

findings for provinces like Savona and Monza and Brianza demonstrate that road length-based indicators are not only more stable but also capable of revealing different facets of road safety risk crucial for tailored policy interventions (Figure 4).

Figure 4 - *Savona road layout by OSM graph. Year 2023.*

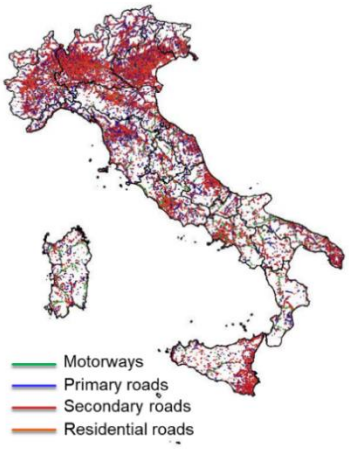


3.6. Towards the Measurement of Traffic Flows

The first result achieved with the measurement of road segment length allows a first step towards correlating road accidents and traffic flows for a more correct measurement of risks. It was essential to start from the knowledge of the length of the national road network by locality to reach the most frequently used indicator of "vehicles per kilometer" per road segments. The refinement process involves identifying road segments with intense traffic flows (using Point of Traffic - PoT data from OSM) and then constructing new synthetic indicators. This ongoing work aims to move beyond static denominators towards dynamic measures of exposure (Figure 5).

Figure 5 visualizes the spatial distribution of these Traffic Points (PoT) across the national territory. The map reveals a significant density of detection points along primary transport corridors, specifically motorways and major state roads and within key metropolitan areas such as Milan, Rome, and Naples. This spatial pattern confirms that the OSM dataset offers substantial coverage of the most intensively used segments of the network. By weighting road segments based on this PoT density, the model can differentiate between high-flow arteries and low-traffic rural roads, providing a crucial correction factor for estimating risk exposure more accurately than simple road length.

Figure 5 - *Point of Traffic - PoT data from OSM. Year 2023.*



4. Conclusions

This study demonstrates the significant potential of leveraging Big Data from OSM for a more nuanced and accurate territorial analysis of road accidents.

Integrating Big Data into official statistics requires continuous methodological adaptation to ensure consistency, quality, and representativeness. It is essential to support innovative data sources with a solid theoretical and statistical foundation to turn raw information into actionable knowledge for public policies. A key methodological contribution here is the innovative application of road carriageway length as a primary denominator, which, unlike traditional measures (resident population, vehicle fleet), is less affected by the distortions caused by human mobility. The inherent stability of the road network's length offers a more robust foundation for assessment. The project using OSM to investigate road accident patterns follows these principles, with ongoing activities expected to lead to further developments. Using non-traditional sources like OSM expands the possibilities for analysing road accidents across space and time. By combining open data, geospatial technologies, and advanced statistical methods, researchers can produce more timely and detailed territorial statistics. This study also shows that OSM-based Big Data can support synthetic indicators that offer clear benefits for policymakers. These tools empower a more evidence-based approach to road safety management, facilitating better-targeted prevention programs. The indicators allow proposing an efficient allocation of resources for infrastructure improvements and effective strategies for addressing unsafe driving practices too. Differences in high-risk province rankings, along with cases like Savona and Monza and Brianza, show the need for indicators that capture real exposure more accurately. Statistical agencies, researchers, and the open-data community must work together to

create shared standards and validation tools, aiming to estimate actual traffic flows on the national road network and to calculate more reliable accident probabilities.
Researchers who apply this methodology in other contexts should adapt the bridge coding table to local classifications, align the chosen OSM snapshot with the accident reference period, and use both advanced GIS skills and statistical rigor to manage the heavy computational work behind national-scale analyses.

References

BROCCOLI M., BRUZZONE S. 2023. Road accidents in Italy: New indicators, at province level, based on geographic information system open data. *Statistics, Technology and Data Science for Economic and Social Development Book of short papers of the ASA Bologna Conference 2023* - Supplement to Vol. 35, No. 3.

BROCCOLI M., BRUZZONE S. 2021. *Use of the open street map to calculate indicators for road accidents on the Italian roads. Updating with 2017 data*. Istat, Rome, Experimental Statistics. https://www.istat.it/en/archivio/257384

BROCCOLI M., BRUZZONE S. 2019. *Use of the open street map to calculate indicators for road accidents on the Italian roads. Year 2016*. Istat, Rome, Experimental Statistics. https://www.istat.it/en/archivio/231740

DE MURO P., MAZZIOTTA M., PARETO A. 2011. Composite Indices of Development and Poverty: an Application to MDGs. *Soc Indic Res*, Vol. 104, pp. 1–18.

Istat 2024. *Road Accidents in Italy. Year 2023*. Press Release.
https://www.istat.it/en/press-release/road-accidents-2023/

HAKKERT A.S., BRAIMAISTER L. 2002: *The uses of exposure and risk in road safety studies*. Number: R-2002-12, SWOV Institute for Road Safety Research, The Netherlands. Leidschendam 2002

ZILSKE M., NEUMANN A., NAGEL K. 2011. Open Street Map for traffic simulation. In *Proceedings of the 1 st European state of the map: OpenStreetMap conference*. - Wien: OpenStreetMap Austria u.a., 2011. - pp. 126–134.

Marco BROCCOLI, Istat - Istituto Nazionale di Statistica broccoli@istat.it
Silvia BRUZZONE, Istat - Istituto Nazionale di Statistica bruzzone@istat.it