RE-ENGINEERING ENVIRONMENTAL DATA COLLECTION IN CITIES[1]

Domenico Adamo, Gianpiero Bianchi, Lucia Mongelli

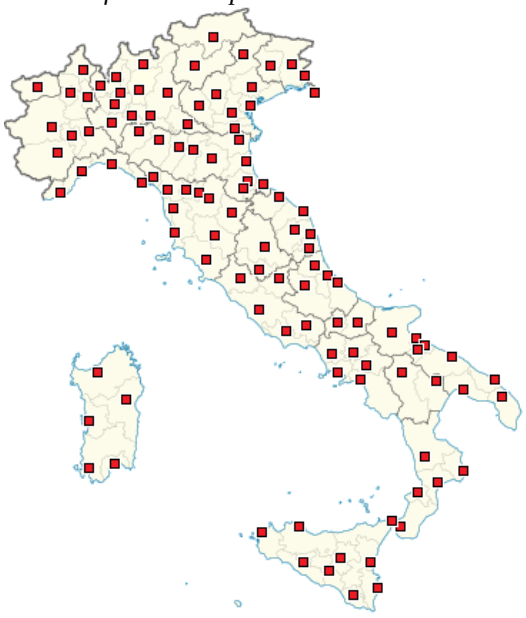
Abstract. The work provides an information framework to support the monitoring of the state of the urban environment and the activities carried out by administrations of provincial capitals to improve quality of the environment in cities. In particular, the "Survey on urban environmental data" is carried out annually by Istat, is included in the National Statistical Program in force and collects environmental information about all Italian capital municipalities. The work describes the study and design of a new validation process, according to a generalized perspective, which includes automatic procedures for checking the consistency of the data collected, monitoring the processing and interaction with the Municipal Statistics Offices. A representation of the rules in formal logic will be adopted, through a metalanguage in order to support an automatic approach. The result is an integrated system of generalized services that works formally and therefore can be used in different contexts. In order to maintain the quality standards of the data disseminated by the survey, a study on the administration of questionnaires on different editions of the survey was designed.

1. A brief overview of the Survey on urban environmental data

Data on the urban environment is a multi-source statistical process, organized in 8 thematic modules: air quality, urban waste, mobility, noise, energy, urban green, water, eco-management, which produces environmental indicators for 110 Italian cities (95 provincial capitals, 14 metropolitan city capitals and the municipality of Cesena, which participates on a voluntary basis), Figure 1.

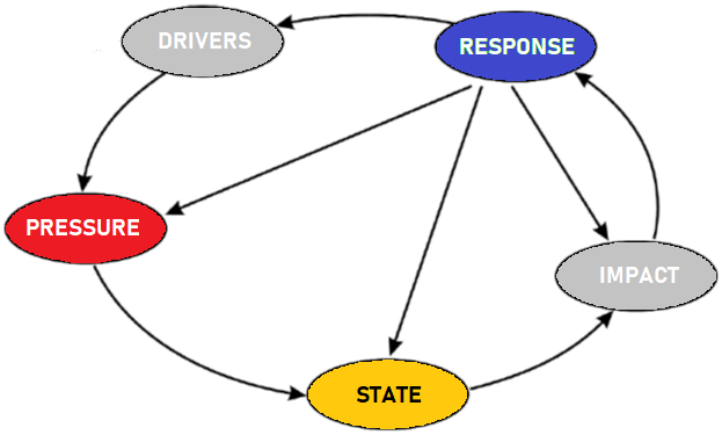
[1]The paper is the result of the common work of the authors. In particular, paragraph 1 is attributed to Domenico Adamo, paragraphs 2 and 3 are attributed to Lucia Mongelli, and paragraphs 4 and 5 are attributed to Gianpiero Bianchi. The conclusions (ph.6) are a joint work of all the authors.

Figure 1 - *Spatial distribution of the municipalities involved in the survey.*



The process provides a comprehensive information framework for monitoring the quality of the urban environment, status and pressure indicators (Adamo et al., 2020), according to the DPSIR model, developed by the European Environment Agency (Fig.2, Bosch et al., 1999) and environmental policies implemented by local governments (so-called response indicators, such as directives, plans, technology development).

Figure 2 – *The DPSIR model.*



The DPSIR model consists of: determinants (agriculture, population, and transport), pressures (waste, emissions), state (water, air), impact (costs, pathologies), responses (directives, policies, technology development); it represents a tool capable of evaluating the causal chain leading to environmental alteration, (measured through environmental indicators).

The urban environment survey is part of the National Statistical Programme (NSP), managed by SISTAN and updated every 3 years. Being included in the NSP as a public interest investigation, data collection is carried out by law and the response is mandatory for reporting units.

The NSP also provides the legal basis for the use of administrative data for statistical purposes. Specific agreements are concluded between Istat and the data controllers to define the characteristics and timing of the provision of data, within the SISTAN regulatory framework and in accordance with the rules for the protection of personal data.

The use of administrative data reduces costs and burden for respondents.

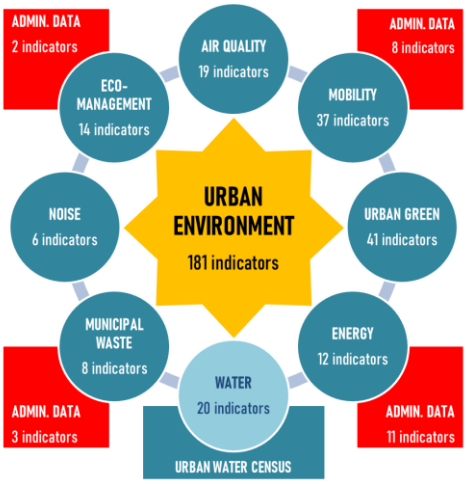
The survey data are collected through thematic questionnaires: Air quality, Eco-management, Noise, Urban waste, Water, Energy, Urban green, Mobility. The theme Water from 2018 is taken from the urban water census, which covers the entire national territory.

The survey data are then supplemented with four modules on particular topics and provided separately by the data controllers.

The whole process handles about 500 elementary variables, to produce the 181 indicators (2020 version), 13% of which are based on administrative data.

Figure 3 summarizes the indicators by theme.

Figure 3 - *Diagram of the data source.*



All indicators are disseminated by municipality, and aggregate estimates are provided by geographical area.

Some indicators are part of the set of statistical measures of Istat for the monitoring of SDGs (Sustainable Development Goals) in Italy, consisting of 17 points, identified by the UN in 2015 that aim to safeguard the planet and the welfare of its inhabitants, with a horizon that reaches up to 2030.

2. Data collection

Since 2008, Istat has introduced important methodological innovations for the "Survey on urban environmental data" with the aim of improving, standardizing the encodings and formats of variables, and simplifying the data collection process.

According to the provisions of the Code of Digital Administration in 2005 (d.lgs 82/2005 and subsequent additions and changes) which provides that data must be transmitted to Istat in computerised mode, the CAWI technique (Computer Assisted Web Interviewing) has been introduced for data acquisition in electronic format, through the Gino++ (Gathering information Online) portal of Istat (Torelli, R. 2011).

GINO++ is a generalized software that allows not only the collection of data but the complete management of surveys via web, creation of web questionnaire, controlled acquisition of data online and/ or file upload, custom site preparation for the survey, monitoring of the status of questionnaires and records, contacts for reminders and reminders, reports.

In addition, on the home page of the Gino Portal (https://gino.istat.it/amburb/) respondents find the support material to fulfill all the obligations provided by the survey: the description of the survey, the detail of the law for the response obligation, instructions for accessing and filling in the questionnaire, IT requirements, FAQs.

The data are collected by the Municipal Statistics Offices, which, through a pre-survey (limesurvay), identify in the Administrations to which they belong a coordinator and one or more persons referencing the survey topics, which are provided with personal credentials to access, enter, modify and save data.

Depending on the topics, the reference persons collect the data directly from the municipalities, or request them from other local authorities (e.g. public transport companies).

Through GINO++ the Municipal Statistics Offices, the coordinators and the referents of the different topics, can send the data through the direct compilation of web questionnaires (CAWI). To improve the completeness and consistency of data entered in the Gino++ data acquisition system, automated checks have been implemented to report anomalies, to prevent inconsistent or invalid or out-of-range data entering and sending questionnaires with missing answers.

An additional monitoring function allows to constantly monitor the activity of respondents, from recording to sending data, including reporting any violations of consistency rules.

3. Process innovations. Validation automation

In line with the objectives of Istat to provide the country with correct statistical information and to innovate the various processes of production of statistical information, consistent with the progressive digitalization of data collection processes, it became necessary to design the use of innovative solutions by re-engineering the validation phase of the questionnaires, with the implementation of additional automatic control rules different from those already provided for by the validity and internal consistency checks of the Gino ++ system.

During this experimental phase, in order to ensure the regular conduct of the survey while maintaining the quality standard of the data collected, a different frequency is expected for some thematic questionnaires, which do not produce indicators intended for institutional dissemination.

For the 2023 edition the thematic of the questionnaire are questionnaires: Air, Mobility, Municipal waste, Noise, Urban green. For the 2024 edition, however, the thematic of the questionnaire will be questionnaires: Air, Eco-management, Energy, Mobility, Urban green.

This innovation could be completed during the two editions of the survey, 2023 and 2024.

The new rules manage, at least in part, the aspects so far entrusted to the review by monitors: in particular the interception of measurement errors, discontinuity of the time series and other anomalous values. These rules would work on a dynamic basis, by comparing the data collected in the current edition with those validated and disseminated in previous editions.

In table 1 the phases of the investigation after the re-engineering.

Table 1 - *Survey stages. After re-engineering.*

		MUNICIPAL OFFICES	STAGES	ISTAT	
				Data collection	Environmental statistics
			Survey design	Survey organization, Implementation of CAWI questionnaires	Information contents and metadata management
Questionnaire states	**Initial** - before taking over by reference person	Registration	**Data collection**	Controlled acquisition through Gino electronic questionnaire with rules Automatic control of: -) measurement errors, -) historical series discontinuities -) other abnormal values Monitoring of survey operations	Assistance to respondents and to data collection staff
	In process - after first opening by reference person	Data entry			
	Sent - after completion by reference person				
	Checked - after preliminary check			Automated return on respondents	
			Data processing		Data editing and validation
			Data dissemination		Data analysis and reporting

4. A generalized data editing for error detection

A generalized editing system allows checking the consistency of the data collected with respect to the check plan for survey data, with intra-record and inter-record rules. Furthermore, the editing system identifies the inconsistency and redundancy in the rules set.

The application classifies exact and incorrect questionnaires, identifying the collected units involved in violated edits together with the fields involved in the violation of the rules.

The application uses a customizable metadata table to apply the editing plan.

This table contains the following information:

- The type of rule: Validity, Logic, Mathematics and Logical-Mathematics;
- The textual description of the rule;
- The representation in formal logic, that is through a meta-language understandable to the editing application;
- Typology of rule between hard (blocker rule) or soft rule (non-blocker rule);

- A hierarchy of rules, to indicate to the application the relationship and the order of control of the rules. In the case of a violated rule, all the related rules (which are a specialization of the rule itself) of a subsequent order can be put directly to violated.

This section provides some useful concepts for the representation of rules in formal logic. In particular, it provides the definition and representation of a list of check rules and for understanding how to transform the textual rules, defined in the examples described below, in compatibility or incompatibility rules when they are translated in a formal language (Bruni and Bianchi, 2012). Rules are expressions typically used to detect, among a possibly large set of elements, the ones verifying some conditions. It is convenient, in order to verify a set of checking rules, to express them using a structure based on propositional logic.

Propositional logic, sometimes called sentential logic, can be considered a grammar for exploring the construction of complex sentences using atomic statements as building blocks connected by logical connectives. In this type of logic, logical formulas (sentences, propositions) are built up from atomic propositions that are unanalysed. The meaning of these atomic propositions will be known for the specific domain of application. A truth assignment to such atomic propositions will determine the truth value of the whole formula according to the truth rules of the logical connectives. The traditional (symbolic) approach to propositional logic is based on a clear separation of the syntactical and semantical functions.

The syntax deals with the laws that govern the construction of logical formulas from the atomic propositions and with the structure of proofs. Semantics, on the other hand, is concerned with the interpretation and meaning associated with the syntactical objects. A basic aspect of propositional calculus is that inferences are obtained as purely syntactic and mechanical transformations of formulas. The set of primary logic connectives $\{\neg, \vee, \wedge\}$, together with the brackets () to distinguish start and end of the field of a logic connective.

- The set of proposition symbols, such as x1, x2, ..., xn.
- The only significant sequences of the above symbols are the well-formed formulas (WFFS). An inductive definition is the following:
- A propositional symbol x or its negation ¬x.
- Other WFFS connected by binary logic connectives ($\vee$, $\wedge$) and surrounded, in case, by brackets.

Both propositional symbols and negated propositional symbols are called literals. Propositional symbols represent atomic (i.e. not divisible) propositions, sometimes called atoms. An example of WFF is the following:

$$(\neg x_1 \vee (x_1 \wedge x_3)) \wedge ((\neg(x_2 \wedge x_1)) \vee x_3) \tag{A.1}$$

A formula is a WFF if and only if there is no conflict in the definition of the fields of the connectives. In order to simplify the exposition, we will henceforth assume that all our formulas are well formed unless otherwise noted.

The calculus of propositional logic can be developed using only the three primary logic connectives above. However, it is often convenient to introduce some additional connectives, such as ⇒ which is called *implies*.

They are essentially abbreviations that have equivalent formulas using only the primary connectives. In fact, if S_1 and S_2 are formulas, we have:

$(S_1 \Rightarrow S_2)$ is equivalent to $(\neg S_1 \vee S_2)$.

The elements of the set {T, F}(or equivalently{1,0}) are called truth values with T denoting True and F denoting False. When all the proposition symbols of a formula receive truth values, the truth or falsehood of that formula is obtained according to the truth rules of the logical connectives (considering their appropriate meaning of "not", "or", and "and"). As an illustration, consider the formula (A.1).

Let us start with an assignment of true (T) for all three atomic propositions x_1, x_2, x_3. At the next level, of sub formulas, we have $\neg x_1$ evaluates to F, $(x_1 \wedge x_3)$ evaluates to T, $(x_2 \wedge x_1)$ evaluates to T, and x_3 is T. The third level has $(\neg x_1 \vee (x_1 \wedge x_3))$ evaluating to T and $((\neg (x_2 \wedge x_1)) \vee x_3)$ also evaluating to T. The entire formula is the "and" of two propositions both of which are true, leading to the conclusion that the formula evaluates to T. This process is simply the inductive application of the rules:

- S is T if and only if ¬S is F.
- $(S_1 \vee S_2)$ is F if and only if both S_1 and S_2 are F.
- $(S_1 \wedge S_2)$ is T if and only if both S_1 and S_2 are T.

Such a truth evaluation approach can be the basis for developing *control rules*, which are rules that allow the individuation of inconsistent or erroneous data records into a large set of similar records. We denote by P a *record schema*, that is a set of *fields* f_i, with $i = 1...m$, and by p a corresponding *record instance*, that is a set of values v_i, one for each of the above fields.

$$P = \{f_1, \dots, f_m\} \quad p = \{v_1,\dots,v_m\} \tag{A.2}$$

Each field f_i, with $i = 1...m$, has its *domain* D_i, which is the set of every possible value for that field. Examples of fields f_i are age or marital status, and corresponding examples of values v_i are 18 or single.

18. A control rule should be applied to a generic record and provide a binary value. Therefore, each rule can be seen as a mathematical function r_k from the Cartesian product of all the domains to the Boolean set {0,1}, as follows (see also Fellegi and Holt, 1976).

$$r_k : \begin{array}{ccc} D_1 \times \ldots \times D_m & \rightarrow & \{0,1\} \\ p & \mapsto & 0,1 \end{array} \qquad \text{(A.3)}$$

The problem of error detection can be approached by formulating a set of rules $R = \{r1, ..., r_t\}$ that are verified by consistent, or correct, records, and are not verified by inconsistent, or erroneous, records. These rules are called compatibility rules, they are such that a generic record p is recognized as a correct record if and only if $r_k(p) = 1$, for all $k = 1, ..., t$. On the other hand, incompatibility rules are verified by erroneous records and not verified by correct records. The detection of erroneous records into a large set of records is a very relevant problem in the field of data E&I.

Compatibility and incompatibility rules can be expressed as disjunction (∨) and/or conjunction (∧) of conditions (also called propositions), hence with the structure of propositional logic formulas. Like to the truth evaluation technique described above, the value of each field of a record under analysis provides a truth assignment for those propositions. The truth/falsehood of the formula constituting the rule provides now the detection of inconsistent or erroneous data records.

However, differently from the case of pure propositional logic, conditions may have an internal structure.

It is necessary to distinguish between two different types of structures for the conditions:

- A condition involving values of a single field is called a logical condition, and corresponds to an atomic proposition of propositional logic. For instance, (*age* < 14) is a logical condition.
- A condition involving mathematical operations between values of fields is called mathematical condition. For instance: (*age - years married* ≥ 14) is a mathematical condition.

We call *logical rules* the rules expressed only with logical conditions, *mathematical rules* the rules expressed only with mathematical conditions, and *logic-mathematical rules* the rules expressed using both types of conditions. For instance, a logical rule expressing that "*if PM10 number of exceedances of the daily average of 50 µg/m3 isn't less than 0, then PM10 annual average concentration value should be not less than 0*" is:

$$PM10\_SUP\_CENTR\_ARIA \geq 0 \Rightarrow PM10\_MEDIA\_CENTR\_ARIA\_T1 \geq 0$$

This rule can be represented by the following compatibility rule:

$$(PM10\_SUP\_CENTR\_ARIA \geq 0) \vee PM10\_MEDIA\_CENTR\_ARIA\_T1 \geq 0$$

or, equivalently, by the following incompatibility rule:

$$PM10\_SUP\_CENTR\_ARIA \geq 0 \wedge \neg (PM10\_MEDIA\_CENTR\_ARIA\_T1 \geq 0).$$

A formal definition of the structure of the rules allows solving by means of automatic formal methods a number of difficult and computationally demanding problems arising in the different steps of E&I procedures. Examples are:

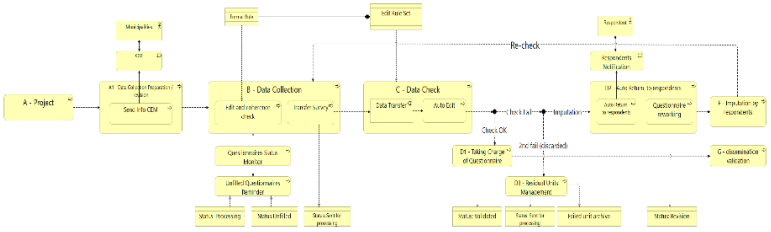
- Problems of error localization (the determination of the erroneous fields of a record).
- Problems of imputation (the determination of the correct values for the erroneous fields of a record. This can be done according to a minimum change principle or by means of a data driven approach).
- Problems of finding contradictions into the set of rules itself (the determination of a (sub)set or rules determining a logical inconsistency).

Note, in particular, that very effective solution approaches are available when encoding rules into linear inequalities. Indeed, a parallelism can be established between logic formulas and linear constraints, and between atomic propositions and 0-1 variables (see Chandru and Hooker, 1991). The above problems are converted into linear or integer linear programming problems and solved by means of efficient optimization solvers. For further details on those techniques see e.g. Bruni and Bianchi, 2012; Bianchi *et al*., 2020; Bianchi *et al*., 2008.

5. New validation process

A sketch of the new validation process is shown in figure 4.

Figure 4 – *Process outline.*



The Process Spans into several Phases

- Survey Project and Preparation (A – A1)
- Data Collection (B)
- Data Check (C - D - E)
- Data transfer to archieves and to production units (G)

A Monitor control the data flow and records questionnaires status and overall completion status

Questionnaires check can fail only once, so unfilled and double failing units are separated and stored elsewhere. After that, validated questionnaires are ready for dissemination.

6. Conclusion

Whereas the ambitious project described above aims to improve data accuracy and consistency through the following progressive design innovations:

1. The design and implementation of a generalized validation process with automated error and analysis reports aimed at increasing the efficiency of the process, increasing the quality of the data collected and reducing the resources employed.
2. The definition of the methodologies and algorithms needed to perform the automatic checks required by the validation process. A main advantage is that this procedure works only at the formal level, so it can be performed without the need of going into the semantic meaning of the validation rules.
3. The design and development of application components and databases, ensuring the integration of the validation process with the acquisition and production environment.
4. The analysis and validation of the implemented tools and the results of the new validation process, through the definition of test cases and continuous experimentation on the 2023 survey, allow the generalised model to be applied to highly differentiated and technically complex situations related to the themes identified by the urban environment survey. For example, for the urban green topic, the test concerns the green management tools used by municipal administrations (qualitative variables); for public transport, on the other hand, the model is applied to the demand and supply of the service (quantitative variables).

The effectiveness in introducing these new generalized solutions that adopt standard methodologies based on technological innovations have the stated goal of applying a new strategy for the pre-validation of data sent by municipalities that by standardizing and automating the recall of incongruents, they want to reduce to a few cases those necessary for in-depth examination by Istat's thematic experts.

References

ADAMO D., COSTANZO L., BUZZI L., LAGANÀ A., GAROZZO S., GRECO V. 2020. *Principali fattori di pressione sull'ambiente nelle città italiane– Anno 2018,* In ADAMO D. and COSTANZO L. (Eds.), Istat, Territori, Letture statistiche.

BIANCHI G., BRUNI R.A. 2012. Formal Procedure for Finding Contradictions into a Set of Rules, *Applied Mathematical Sciences*, Vol. 6, No. 126, pp. 6253-6271.

BIANCHI G., MANZARI A., REALE A. 2008. An overview of editing and imputation methods for the next Italian censuses. In *Proceedings of the Conference of European statisticians,* Geneva.

BIANCHI G., MANZARI A., PEZONE A., REALE A., SAPORITO G. 2005. New procedures for editing and imputation of demographic variables, In *Proceedings of the Conference of European statisticians,* Ottawa.

BOSCH P., BÜCHELE M., GEE D. 1999. Environmental Indicators: Typology and Overview, *European Environment Agency*, pp. 19.

CHANDRU V., HOOKER J.N. 1991. Extended Horn Sets in Propositional Logic, *J. ACM,* Vol. 38, pp. 205-221.

FELLEGI I.P., HOLT I.P.D. 1976. A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, Vol. 71, pp. 17-35.

TORELLI R. 2011. A generalized system for web surveys, In *Proceedings of Statistics Canada Symposium.*

Domenico ADAMO, Istat, adamo@istat.it
Gianpiero BIANCHI, Istat, gianbia@istat.it
Lucia MONGELLI, Istat, mongelli@istat.it