Environmental Risk and Public Procurement in Italy

Thomas Bassetti*

In this document, we illustrate how the dataset titled "Environmental Risk and Public Procurement" (ERPP) was constructed and describe the variables included in it.

1. Data Collection and Cleaning

The ERPP dataset is based on two sources of information. Public procurement data are collected by the Autorità Nazionale Anticorruzione (ANAC), whereas hydrogeological risk assessments are collected by the Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA).

ANAC (2025, March 6) provides a large open database on public procurement contracts. In general, two datasets (namely Anagrafica and CUP) are updated on a monthly basis and are displayed in a "FULL" format. For the remaining datasets, ANAC switched to a new format called "DELTA" after March 2023. Based on an incremental strategy, this format updates CIGs only when changes are made. DELTA sources have the same names as FULL sources with the addition of the yyyymmdd prefix.

Among the available files, we considered four categories of data: CIG (Codice Identificativo di Gara, i.e., auction code), SMART CIG (for special contracts below 40,000 euros), Aggiudicazioni (adjudications), and Categorie-opera (work categories). For each category, we created a single dataset that included all datasets in both FULL and DELTA formats and was cleansed of duplicates. The operations were all performed using Stata 18. We describe the steps involved in collecting and cleaning data below.

1.1 Creating CIGs+SMARTCIGs

As a first step, we merged the CIG and SMARTCIG datasets, obtaining 25,095,780 observations.[1] Due to the size of the new dataset, we kept only the information relevant to our purposes. More specifically, we kept the following variables: the auction code (CIG), the auction value in euros, the object (as a service, work, or equipment), the ISTAT municipality code, the publication date, and the offer deadline.

* Department of Economics and Management "Marco Fanno", University of Padua (Italy), Via del Santo 33, 35123, Padua, e-mail: thomas.bassetti@unipd.it

[1] Data from the CIG dataset was filtered to exclude auctions that were canceled, deserted, or had inadmissible bids.

1.2 Merging Categoria-opera with CIGs+SMARTCIGs

Our next step was to merge the CIGs+SMARTCIGs dataset with the category-opera dataset, after removing duplicates and categorizing each auction according to its category (id_categoria). Because the final dataset should focus on flood risks, we have divided public expenditures into green and gray categories. Public expenditures devoted to mitigating hydrogeological risk are referred to as green expenditures, while all other expenditures are referred to as gray expenditures. Unfortunately, for 69,700 observations out of 25,095,780 (i.e., 0.28% of observations), it was not possible to identify a spending category.

Table 1 lists the categories that are considered to be green.

Table 1. Green categories

Italian	English
OG 4 - OPERE D'ARTE NEL SOTTOSUOLO	OG 4 - UNDERGROUND WORKS
OG 5 - DIGHE	OG 5 - DAMS
OG 6 - ACQUEDOTTI, GASDOTTI, OLEODOTTI, OPERE DI IRRIGAZIONE E DI EVACUAZIONE	OG 6 - AQUEDUCTS, GAS PIPELINES, OIL PIPELINES, IRRIGATION AND EVACUATION WORKS
OG 7 - OPERE MARITTIME E LAVORI DI DRAGAGGIO	OG 7 - MARITIME WORKS AND DREDGING WORKS
OG 8 - OPERE FLUVIALI, DI DIFESA, DI SISTEMAZIONE IDRAULICA E DI BONIFICA	OG 8 - RIVER WORKS, DEFENSE, HYDRAULIC SYSTEMS AND DRAINAGE WORKS
OG12 - OPERE ED IMPIANTI DI BONIFICA E PROTEZIONE AMBIENTALE	OG12 - DRAINAGE AND ENVIRONMENTAL PROTECTION WORKS AND SYSTEMS
OG13 - OPERE DI INGEGNERIA NATURALISTICA	OG13 - NATURALISTIC ENGINEERING WORKS

1.3 Public spending indicators

The merged dataset described in point 1.2 allows us to calculate four aggregate indicators at the municipal level. Two indicators are used to identify the type of public expenditure: green versus gray. In particular, we aggregated green expenditure and gray expenditure to obtain total expenditure in green projects (tot_green) and total expenditure in gray projects (tot_gray).

There are two other indicators to be considered, which refer to the quality of public spending and, in particular, the possibility that public projects are developed to favor some private parties rather than ensure the best price-quality ratio. Indeed, the Italian legislation permits local authorities to assign public works to private contractors without requiring a public audit procedure when the contract value is below a certain level. This possibility may facilitate corruption by preventing competition among agents. To assess corruption risk at the municipal level, we constructed a variable that measures the concentration of public expenditure below the threshold values defined by Italian legislation for the direct assignment of public contracts. This measure is the ratio of public spending below and above

the threshold. Thresholds have changed over time and vary based on the main object of the contract, such as "lavori" (work) or "servizi e forniture" (services and equipment).[2]

Table 2 shows how these thresholds have changed over time.

Table 2. Italian thresholds for direct assignments

Year	Service & Equipment	Works	Legislation
Before 2020	$\leq 40{,}000€$	$\leq 40{,}000€$	D.lgs. 163/2006
2020-2023	$\leq 75{,}000€$	$\leq 150{,}000€$	D.L. 76/2020 (L. 120/2020)
Since 2023	$\leq 140{,}000€$	$\leq 150{,}000€$	D.Lgs. 36/2023

Note: The D.Lgs. 36/2023 became effective since July 1st, 2023; although, many norms co-existed with the new ones till December 31st, 2023.

Following ANAC, we exclude contracts that are too far from the threshold to be considered anomalous. Table 3 provides the thresholds for measuring concentration ratios: the total value of auctions within the thresholds and the remaining value of auctions outside of the thresholds.

Table 3. Thresholds used to measure concentration ratios

Year	Service & Equipment	Works
Before 2020	$20{,}000 < value \leq 40{,}000€$	$20{,}000 < value \leq 40{,}000€$
2020-2023	$37{,}500 < value \leq 75{,}000€$	$75{,}000 < value \leq 150{,}000€$
Since 2023	$69{,}500 < value \leq 139{,}000€$	$75{,}000 < value \leq 150{,}000€$

Source: ANAC (2024, June 21).

ANAC provides a similar concentration index. This index of concentration is calculated by dividing the number of contracts below the threshold by the number of contracts above it. Nevertheless, this indicator is not sensitive to the fact that a municipality with a higher level of corruption should display a greater concentration of values near the threshold than a municipality with a lower level of corruption. In contrast, we consider the relative amount of public expenditure below the threshold as a better indicator of possible corrupt behavior. Finally, we have defined two separate indicators for "services & equipments" and "works", improving the information set. In Section 3, we validate our concentration ratios.

1.4 Using Aggiudicazioni to identify anomalous rebates

After having removed the oldest duplicates from the Aggiudicazioni dataset, we merged it with the dataset described above. To simplify data management, we kept only the information that is relevant to us. Specifically, we kept the following variables: cig (auction code), massimo_ribasso (maximum rebate), minimo_ribasso (minimum rebate), importo_aggiudicazione (final value),

[2] Additionally, we include in the numerator those contracts assigned directly under other legal circumstances, even if their value is above the threshold.

ribasso_aggiudicazione (winning rebate), numero_offerte_ammesse (number of admitted offers), numero_offerte_escluse (number of excluded offers), id_aggiudicazione (ID of the winner).

These variables were used to create a dummy variable that takes the value 1 if the maximum rebate is more than 20% higher than the mean value determined by averaging the minimum and maximum rebates. The use of this indicator is due to the lack of information about the entire distribution of rebates; however, it has the advantage of minimizing the probability of a false alarm (Conti and Naldi, 2008).[3] By averaging this dummy at the municipal level, we obtained the variable *anomalies.* This variable represents the annual proportion of audits in which the maximum rebate was significantly low. In Section 3, we validate the variable.

We collapse the sample at the municipal level, keeping the following variables: istat_code, year, concentration_works, concentration_serv, tot_green, tot_gray, mean_anomalies.

1.5 Merging with ISPRA data

Our final step was to merge the previous dataset with ISPRA data using the istat_code. Overall, we collected 119,372 observations across 7,903 municipalities between 2007 and 2024.

ISPRA data refer to two types of risk: the flood risk and the landslide risk. These measures of risk vary across municipalities but remain constant over time. The term "flood" refers to water temporarily covering areas that are normally dry. Following Italian legislation (D.Lgs. 49/2010), the flood risk has been divided into three classes of risk: low risk (P1, when the flood occurs rarely, every 200 years on average), medium risk (P2, when the flood occurs every 100 years on average), and high risk (P3, when the flood occurs frequently, every 20-50 years on average).

Regarding the landslide risk, ISPRA identifies five different classes of landslide risk: P4 (Very high landslide risk area), P3 (High landslide risk area), P2 (Medium landslide risk area), P1 (Low landslide risk area), and AA (areas under attention).

Flood and landslide risk classifications refer to different assets at risk, including areas, populations, families, buildings, firms, and cultural heritages. The indicators are expressed in specific units of measurement and as percentages of the municipal endowment.

2. List of variables

Table 4 provides the list of variables that can be found in the ERPP dataset and the corresponding labels.

[2] Data on all participants in public auctions was not available at the time the data was created. Therefore, the current information set limits the possibility of classifying anomalous bids based on rebate proposals from all bidders.

Table 4. List of variables and corresponding labels

Variable	Label
cod_reg	Region code
cod_prov	Province code
pro_com	Municipality code
municipality	Municipality name
ar_kmq	Area (km-sq)
ar_id_p3	High flood risk areas (km-sq)
ar_id_p2	Medium flood risk areas (km-sq)
ar_id_p1	Low flood risk areas (km-sq)
aridp3_p	High flood risk areas (%)
aridp2_p	Medium flood risk areas (%)
aridp1_p	Low flood risk areas (%)
pop_res011	Residents - Census 2011 (n.)
pop_gio	Young population at risk
pop_gio_p	Young population at risk (%)
pop_adu	Adult population at risk
pop_adu_p	Adult population at risk (%)
pop_anz	Old population at risk
pop_anz_p	Old population at risk (%)
pop_idr_p3	Population in high flood risk areas (n.)
pop_idr_p2	Population in medium flood risk areas (n.)
pop_idr_p1	Population in low flood risk areas (n.)
popidp3_p	Population in high flood risk areas (%)
popidp2_p	Population in medium flood risk areas (%)
popidp1_p	Population in low flood risk areas (%)
fam_tot	Families - Census 2011 (n.)
fam_idr_p3	Families in high flood risk areas (n.)
fam_idr_p2	Families in medium flood risk areas (n.)
fam_idr_p1	Families in low flood risk areas (n.)
famidp3_p	Families in high flood risk areas (%)
famidp2_p	Families in medium flood risk areas (%)
famidp1_p	Families in low flood risk areas (%)
ed_tot	Buildings - Census 2011 (n.)
ed_idr_p3	Buildings in high flood risk areas (n.)
ed_idr_p2	Buildings in medium flood risk areas (n.)
ed_idr_p1	Buildings in low flood risk areas (n.)
edidp3_p	Buildings in high flood risk areas (%)
edidp2_p	Buildings in medium flood risk areas (%)
edidp1_p	Buildings in low flood risk areas (%)
im_tot	Firm local units - Census 2011 (n.)
im_idr_p3	Firm local units in high flood risk areas (n.)
im_idr_p2	Firm local units in medium flood risk areas (n.)
im_idr_p1	Firm local units in low flood risk areas (n.)

imidp3_p	Firm local units in high flood risk areas (%)
imidp2_p	Firm local units in medium flood risk areas (%)
imidp1_p	Firm local units in low flood risk areas (%)
n_vir	Cultural heritage - Vincoli in Rete VIR (n.)
bbcc_id_p3	Cultural heritage in high flood risk areas (n.)
bbcc_id_p2	Cultural heritage in medium flood risk areas (n.)
bbcc_id_p1	Cultural heritage in low flood risk areas (n.)
bbccidp3_p	Cultural heritage in high flood risk areas (%)
bbccidp2_p	Cultural heritage in medium flood risk areas (%)
bbccidp1_p	Cultural heritage in low flood risk areas (%)
ar_fr_p4	Very high landslide risk area P4 (km-sq)
ar_fr_p3	High landslide risk area P3 (km-sq)
ar_fr_p2	Medium landslide risk area P2 (km-sq)
ar_fr_p1	Low landslide risk area P1 (km-sq)
ar_fr_aa	Landslide risk area under attention AA (km-sq)
ar_fr_p3p4	Very high and high landslide risk area (km-sq)
ar_frp4_p	Very high landslide risk area P4 (%)
ar_frp3_p	High landslide risk area P3 (%)
ar_frp2_p	Medium landslide risk area P2 (%)
ar_frp1_p	Low landslide risk area P1 (%)
ar_fraa_p	Landslide risk area under attention AA (%)
ar_frp3p4p	Very high and high landslide risk area (P4 + P3) (%)
pop_fr_p4	Population in very high landslide risk area P4 (n.)
pop_fr_p3	Population in high landslide risk area P3 (n.)
pop_fr_p2	Population in medium landslide risk area P2 (n.)
pop_fr_p1	Population in low landslide risk area P1 (n.)
pop_fr_aa	Population in areas under attention AA (n.)
popfr_p3p4	Population in high and very high landslide risk area (P4 + P3) (n.)
popfrp4_p	Population in very high landslide risk area P4 (%)
popfrp3_p	Population in high landslide risk area P3 (%)
popfrp2_p	Population in medium landslide risk area P2 (%)
popfrp1_p	Population in low landslide risk area P1 (%)
popfraa_p	Population in areas under attention AA (%)
popfrp3p4p	Population in high and very high landslide risk area (P4 + P3) (%)
fam_fr_p4	Families in very high landslide risk area P4 (n.)
fam_fr_p3	Families in high landslide risk area P3 (n.)
fam_fr_p2	Families in medium landslide risk area P2 (n.)
fam_fr_p1	Families in low landslide risk area P1 (n.)
fam_fr_aa	Families in areas under attention AA (n.)
famfr_p3p4	Families in high and very high landslide risk area (P4 + P3) (n.)
famfrp4_p	Families in very high landslide risk area P4 (%)
famfrp3_p	Families in high landslide risk area P3 (%)
famfrp2_p	Families in medium landslide risk area P2 (%)
famfrp1_p	Families in low landslide risk area P1 (%)

famfraa_p	Families in areas under attention AA (%)
famfrp3p4p	Families in high and very high landslide risk area (P4 + P3) (%)
ed_fr_p4	Buildings in very high landslide risk area P4 (n.)
ed_fr_p3	Buildings in high landslide risk area P3 (n.)
ed_fr_p2	Buildings in medium landslide risk area P2 (n.)
ed_fr_p1	Buildings in low landslide risk area P1 (n.)
ed_fr_aa	Buildings in areas under attention AA (n.)
ed_fr_p3p4	Buildings in high and very high landslide risk area (P4 + P3) (n.)
edfrp4_p	Buildings in very high landslide risk area P4 (%)
edfrp3_p	Buildings in high landslide risk area P3 (%)
edfrp2_p	Buildings in medium landslide risk area P2 (%)
edfrp1_p	Buildings in low landslide risk area P1 (%)
edfraa_p	Buildings in areas under attention AA (%)
edfrp3p4p	Buildings in high and very high landslide risk area (P4 + P3) (%)
im_fr_p4	Firm local units very high landslide risk P4 (n.)
im_fr_p3	Firm local units high landslide risk P3 (n.)
im_fr_p2	Firm local units medium landslide risk P2 (n.)
im_fr_p1	Firm local units low landslide risk P1 (n.)
im_fr_aa	Firm local units in areas under attention AA (n.)
imfr_p3p4	Firm local units high and very high landslide risk(P4 + P3) (n.)
imfrp4_p	Firm local units very high landslide risk P4 (%)
imfrp3_p	Firm local units high landslide risk P3 (%)
imfrp2_p	Firm local units medium landslide risk P2 (%)
imfrp1_p	Firm local units low landslide risk P1 (%)
imfraa_p	Firm local units in areas under attention AA (%)
imfrp3p4p	Firm local units high and very high landslide risk(P4 + P3) (%)
bbcc_fr_p4	Cultural heritage very high landslide risk P4 (n.)
bbcc_fr_p3	Cultural heritage high landslide risk P3 (n.)
bbcc_fr_p2	Cultural heritage medium landslide risk P2 (n.)
bbcc_fr_p1	Cultural heritage low landslide risk P1 (n.)
bbcc_fr_aa	Cultural heritage in areas under attention AA (n.)
bbccfrp3p4	Cultural heritage high and very high landslide risk(P4 + P3) (n.)
bbccfrp4_p	Cultural heritage very high landslide risk P4 (%)
bbccfrp3_p	Cultural heritage high landslide risk P3 (%)
bbccfrp2_p	Cultural heritage medium landslide risk P2 (%)
bbccfrp1_p	Cultural heritage low landslide risk P1 (%)
bbccfraa_p	Cultural heritage in areas under attention AA (%)
bbccfrp34p	Cultural heritage high and very high landslide risk(P4 + P3) (%)
istat_code	ISTAT municipality code
year	Year of auction publication
concentration_works	Concentration of offers (in €) below the work threshold for direct assignment
concentration_serv	Concentration of offers (in €) below the service threshold for direct assignment
tot_green	Total public expenditure in green (flood protection) projects
tot_gray	Total public expenditure in other projects

mean_anomalies	Municipal average of auctions with anomalous bids

3. Validation of the anomaly indicator and concentration ratios

Sections 1.3 and 1.4 describe the construction of three variables devoted to measuring the prevalence of corruption at the municipal level. Two indicators, concentration_works and concentration_serv, aim to measure the tendency at the municipal level to publish auctions with a value that falls just below the maximum threshold for direct assignment. These variables should reflect the authorities' willingness to favor certain private contractors and avoid more competitive bidding processes. A third indicator pertains to the behavior of private agents during public auctions and reflects the frequency of observing anomalous bids.

To validate these variables, we used an external dataset containing three indicators that are supposed to be related to corruption. The dataset in question comes from ANAC and covers 745 municipalities with a population of 15,000 or more between 2015 and 2022. The first indicator is a dummy taking the value of 1 if the municipality has been affected by a mafia dissolution order and zero otherwise. It is, however, impossible to use this variable as a validation variable because it exhibits too little variability. Indeed, only 72 out of 5,915 observations take the value of 1 (i.e., 1.22% of the sample). As a result, we take advantage of the other two variables: the risk of contagion and the sub-threshold thickening. Risk of contagion refers to the percentage of municipalities belonging to the same province as the focal municipality and experiencing corruption episodes in the reference year. Because corruption is a "contagious" phenomenon, the presence of municipalities with cases of corruption in the province may contribute to determining whether or not the province is at risk of corruption. Sub-threshold thickening is calculated as the ratio between the number of contracts with an auction base amount between €20,000 and €39,999 and the number of contracts over €40,000.

Table 5 reports the main descriptive statistics for the variables involved in our validation experiment.

Table 5. Descriptive statistics

Variable	Obs.	Mean	Std. dev.	Min	I qrt.	Median	III qrt.	Max
Risk of contagion	1868	0.145	0.126	0.020	0.070	0.120	0.170	1
Sub-threshold thickening	5865	2.710	3.906	0	1.125	1.737	2.833	74
mean_anomalies	5746	0.513	0.252	0	0.333	0.500	0.667	1
concentration_works	5826	0.167	0.394	0	0.009	0.061	0.182	9.954
concentration_serv	5887	0.300	0.467	0	0.073	0.197	0.405	21.764

Table 6 shows the OLS coefficients of different linear models that regress the risk of contagion on the other indicators of potential corruption. While sub-threshold thickening is negatively correlated with contagion risk, our three indicators exhibit a positive, statistically significant correlation with contagion risk. In particular, mean_anomalies and concentration_serv are always positively correlated with contagion risk, whereas, due to collinearity problems, the coefficient of

concentration_works becomes statistically insignificant if considered together with concentration_serv.

Table 6. Risk of contagion (OLS)

	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.150***	0.127***	0.142***	0.133***	0.115***	0.086***
	(0.003)	(0.005)	(0.003)	(0.003)	(0.006)	(0.009)
Sub-threshold thickening	-0.001***				-0.001***	-0.001***
	(0.000)				(0.000)	(0.000)
mean_anomalies		0.034***			0.035***	0.041***
		(0.009)			(0.009)	(0.009)
concentration_works			0.041**		0.040**	0.020
			(0.016)		(0.020)	(0.019)
concentration_serv				0.061***	0.058***	0.056***
				(0.013)	(0.014)	(0.014)
year=2016						0.005
						(0.009)
year=2017						0.036***
						(0.010)
year=2018						0.045***
						(0.010)
year=2019						0.050***
						(0.009)
Observations	1852	1770	1829	1856	1732	1732
Adjusted R^2	0.002	0.005	0.003	0.012	0.020	0.044
F-statistics	13.357	14.707	6.302	20.374	10.567	10.273
Model's DF	1	1	1	1	4	8
Model's p-value	0.000	0.000	0.012	0.000	0.000	0.000

Notes. Standard errors in parentheses. Robust standard errors in parentheses. Significance levels: *p<0.1, **p<0.05, ***p<0.01.

To conclude the validation procedure, we present in Table 7 both the standardized beta coefficients and the dominance analysis based on Column 5 of Table 6. A standard beta coefficient represents the change in the standard deviations of the dependent variable as a result of an increase in a covariate by one standard deviation. The dominance coefficients illustrate the contribution of each covariate to the overall model fit statistic. According to these results, concentration_serv is the best predictor, followed by mean_anomalies. The results of these tests support the use of our risk indicators.

Table 7. Standardized beta coefficients and dominance

	Beta coefficients	Standardized dominance	Ranking
	(1)	(2)	(3)
Sub-threshold thickening	-0.041	0.0823	4
mean_anomalies	0.075	0.248	2
concentration_works	0.053	0.1419	3
concentration_serv	0.104	0.5278	1

References

ANAC (2024, June 21) Nota Metodologica-Aggiornamento maggio 2024 - Indicatori Comunali, https://www.anticorruzione.it/en/rischio-a-livello-comunale (last accessed on March 7th, 2025)

ANAC (2025, March 6). Portale dei dati aperti dell'Autorità Nazionale Anticorruzione https://dati.anticorruzione.it/opendata/dataset.

Conti, P. L., & Naldi, M. (2008). Detection of anomalous bids in procurement auctions. *Decision Support Systems*, *46*(1), 420-428.