

Identificazione di dati di prospettiva delle popolazioni europee nelle indagini UE

Antonella Nannavecchia^{1*} - Francesco D. d'Ovidio² - Silvia d'Ovidio³

¹LUM University – Casamassima Bari (Italy)

²University of Bari Aldo Moro, Bari, Italy – CEDIMES Puglia

³MD «European Affairs», University Sciences Po, Paris - France

Abstract:

Very large amounts of data are generated by the interaction of people with social networks, smartphones, GPS, etc., creating the need for specific technologies (Internet of Things, Big Data Analysis and so on). Such data and tools are key instruments for transforming traditional cities into *Smart Cities*, as well as for regional development. At the international level, six operational areas are recognised for development: smart economy, people, governance, mobility, living, environment. A model that considers and combines these operational areas can be used to better understand the *smartness* level of a city, and to monitor its performance for assessing progress.

European governments, like those of the whole world, are familiar with the importance of these tools to organise urban and extra-urban services in the light of the needs of citizens, their mobility, the environment; in a nutshell, all the elements that form the operational areas of smart cities. In addition to experimenting with algorithms for collecting and using data automatically provided by the devices that people use in their daily lives, information is also gathered about the needs and opinions of citizens.

The latter, unlike any information of "technological" origin, can only be obtained through sample surveys, however extensive, and its use combined with technological data requires some precautions and preliminary operations: for starters, to identify the subjective information that interacts the most with objective data, and therefore useful for their optimal use.

In this article, data from the European Social Survey 2017 are used to test a useful methodology. The European Social Survey (ESS) is an academically driven cross-national survey, whose aims are:

- to monitor and interpret changing public attitudes and values within Europe and to investigate how they interact with Europe's ever-changing institutions,
- to advance and consolidate improved methods of cross-national survey measurement in Europe and beyond,
- to develop a series of European social indicators, including attitudinal criteria.

In the 2017 edition, Round 8, the survey covered 23 countries (largely EU countries, along with Switzerland, Israel and the Russian Federation) employing the most rigorous methodologies in order to collect several information and opinions (more than 400 variables) from circa 44.400 European citizens.

The key issue is to identify, in this stream of information, the variables that are actually important for joint analyses with other data (for example, regional economic indicators), considering that methods of statistical inference are not useful with Large and Big Data. Therefore, we chose to test multivariate statistical methods such as Categorical Principal Component Analysis (CatPCA) and Artificial Neural Network Analysis, particularly Self-Organizing Maps.

Keywords: Smart Cities; European Social Survey; CatPCA; Artificial Neural Network Analysis.

JEL classification: C38, D03, R38, R58

* Corresponding Author: nannavecchia.a@gmail.com .

Il lavoro, frutto di un progetto comune, nelle sue componenti è attribuibile come segue: S. d'Ovidio ha sintetizzato i concetti introduttivi del § 1, A. Nannavecchia ha curato la ricerca e le analisi bivariate, elaborando i paragrafi 2, 4 e 5, F. d'Ovidio ha formulato l'analisi descritta nel § 3 e le conclusioni (§ 6).

1. Introduzione

“Smart City” è un concetto ampio, legato a uno sviluppo sociale ed economico sostenuto da tecnologie basate sull’utilizzo di sensori, Big Data, Open Data e nuove modalità di connessione e scambio di informazioni (cfr. Gretzel et al., 2015). L’ICT e le tecnologie emergenti come l’IoT (*Internet of Things*) e l’IoE (*Internet of Everything*) sono fattori abilitanti per trasformare le città in Smart City e favorire lo sviluppo territoriale (Barletta et al., 2020). L’utilizzo di algoritmi di *machine learning* e intelligenza artificiale può rendere le amministrazioni più efficienti e consentire di offrire servizi migliori in tutti gli ambiti come la gestione del traffico, il controllo dell’inquinamento, trasporti pubblici, ambiente ecc. (cfr. Soomro et al., 2018; Barletta et al., 2019).

Il concetto di Smart City si sviluppa sulla base di sei dimensioni riconosciute a livello internazionale: *smart people*, *smart governance*, *smart economy*, *smart mobility*, *smart environment* e *smart living* (cfr. Anthopoulos et al., 2019). Un modello che prende in considerazione tali dimensioni può essere utilizzato per comprendere meglio il livello di *smartness* di una città e monitorarne i progressi.

Tuttavia, sebbene queste sei dimensioni siano ampiamente riconosciute, il concetto di *smartness* può assumere caratteristiche diverse a seconda delle esperienze locali e dei contesti specifici.

Le città europee, come quelle nel resto del mondo, stanno compiendo enormi sforzi per diventare più innovative e più *smart*. Tuttavia, ogni città ha un tessuto economico ed un contesto politico e sociale diverso, oltre ad essere caratterizzata da condizioni sensibilmente differenti in termini demografici, economici, geografici (cfr. Alawadhi et al., 2012). Le attività intraprese in ambito di Smart City riflettono tali differenze di contesti e condizioni. Ogni città si contraddistingue per un contesto unico in cui sviluppa iniziative innovative e può progettare in maniera unica la propria strategia *smart*. Il successo delle iniziative nel settore Smart City dipende dalla capacità di intraprendere in maniera differenziata e contestualizzata le relazioni tra gli attori chiave e l’ambiente politico ed economico (Odendaal, 2003). Le politiche urbane sono, infatti, strettamente collegate e influenzate dalle peculiarità ambientali, sociali, economiche, culturali, politiche e demografiche. Non esiste, quindi, un unico approccio per raggiungere l’innovazione, ma è opportuno tener conto del mutevole contesto geopolitico (cfr. Eger, 2009). I governi devono, pertanto, comprendere le condizioni ambientali in cui operano e sviluppare un insieme di strategie chiare e adeguate.

Sebbene gli aspetti tecnologici siano quelli più evidenti in ambito Smart City, questi non ne costituiscono l’unica componente. I fattori essenziali che caratterizzano iniziative Smart City, infatti, si distinguono in tecnologici, manageriali e politici. Le varie definizioni di Smart City mettono in rilievo come componente essenziale dello sviluppo urbano anche il fattore umano. *Smart people* è una componente fondamentale che comprende vari fattori quali *life-long learning*, flessibilità, creatività, pluralità etniche e sociali, inclusione sociale, partecipazione alla vita pubblica, infrastrutture sociali, diversità culturale. Il capitale sociale e intellettuale è, pertanto, una risorsa fondamentale per una Smart City e costituisce un elemento trainante in ottica *smart* (Nam & Pardo, 2011).

2. Materiali e metodi

Nel presente lavoro sono stati presi in considerazione i dati della European Social Survey (ESS) 2017, un’indagine condotta con cadenza biennale a livello accademico in varie nazioni dallo *European Research Infrastructure Consortium* (ESS ERIC). Le finalità dell’indagine sono: misurare i cambiamenti nelle attitudini pubbliche e nei pattern comportamentali nel corso del

tempo; migliorare la qualità delle misurazioni comparative quantitative; definire indicatori attitudinali robusti per affiancare gli indicatori comportamentali del benessere nazionale esistenti. Nell'ottava edizione, l'indagine ha riguardato 23 paesi, in maggioranza Stati UE, con l'aggiunta di Svizzera, Israele e Federazione Russa. Sono state rilevate, tramite un'articolata scheda-questionario, numerose informazioni e opinioni (oltre 500 variabili) in ambito politico, economico, ambientale, demografico, welfare, ecc. in un contesto di cambiamento a livello europeo.

Il presente lavoro prende in analisi le opinioni dei cittadini considerando le variabili attinenti ai sei ambiti Smart City (*people, governance, economy, mobility, environment e living*) con l'obiettivo di comprendere come si caratterizzano a livello territoriale. Si tratta essenzialmente di un'analisi preliminare delle variabili da utilizzare in una successiva ricerca che prenderà in considerazione le informazioni più discriminanti a livello territoriale e un set di indicatori quantitativi prodotti da Eurostat in ambito di *Digital Economy and Society*.

L'indagine ESS a cui si fa qui riferimento raccoglie le informazioni fornite da circa 44.400 cittadini europei di età superiore ai 14 anni residenti in abitazioni private, indipendentemente dalla loro nazionalità o cittadinanza, nei seguenti Paesi partecipanti:

- *Paesi appartenenti all'Unione Europea*: Austria, Belgio, Estonia, Finlandia, Francia, Germania, Irlanda, Italia, Lituania, Paesi Bassi, Polonia, Portogallo, Regno Unito, Repubblica Ceca, Slovenia, Spagna, Svezia, Ungheria;
- *Paesi non appartenenti all'Unione Europea*: Federazione Russa, Islanda, Israele, Norvegia, Svizzera.

Nella prima fase è stato studiato il dataset ESS contenente 534 variabili, divise negli 11 gruppi riportati in Tab. 1:

Tabella 1. Raggruppamento variabili indagine European Social Survey (ESS) 2017

Gruppo	N. variabili
Paese	1
Pesi	3
Fiducia nella società e nei media	6
Politica	91
Benessere individuale, esclusione sociale, religione, identità nazionale ed etnica	75
Cambiamento climatico	32
Welfare	42
Genere, anno di nascita e relazioni familiari	37
Caratteristiche sociodemografiche	210
Valori umani	21
Variabili amministrative	16
<i>Totale variabili</i>	<i>534</i>

Lo scopo ultimo del presente studio è valutare le differenze di opinione degli intervistati residenti nei diversi Paesi sugli argomenti attinenti al concetto di Smart Cities. Risulta dunque opportuno identificare quali variabili del dataset siano maggiormente discriminanti, ossia quelle che caratterizzano in misura rilevante i cittadini dell'uno o dell'altro Paese.

Delle 534 variabili sono state inizialmente selezionate quelle strettamente attinenti ai sei ambiti Smart City: 215, circa il 40% del totale. Sono state poi eliminate tutte le variabili che presentassero una quota di dati mancanti (comprendendo anche i rifiuti di fornire l'informazione e le risposte non fornite in quanto argomento ignoto o "non attinente") superiore al 50% del totale dei partecipanti all'indagine.

Di seguito è stata effettuata una analisi bivariata finalizzata a valutare l'associazione tra le variabili restanti, ancora numerose, e la variabile di localizzazione *Country* (residenza). Tale

analisi, dovendosi condurre su variabili nominali, ordinali e quantitative (pur riducibili in intervalli) è stata condotta mediante utilizzo dell'indice *V di Cramer*, rimuovendo dalla lista tutte le variabili con indice inferiore a 0,15. Da questa analisi, sono risultate statisticamente rilevanti ben 48 variabili, di cui la maggior parte misurate su scala ordinale, alcune su scala nominale, poche su scala di intervalli. A questo punto, si è ritenuta necessaria un'ulteriore esplorazione dei dati, utilizzando questa volta una metodologia di analisi multivariata.

3. Selezione finale delle variabili di studio

Tra le metodologie di analisi multivariata, si è ritenuta adeguata una tecnica di eliminazione *backward stepwise* basata sui risultati di una analisi di componenti principali categoriche (CATPCA, acronimo appunto di CATegorical Principal Component Analysis, un algoritmo non parametrico ALSOS-Alternative Least Squares Optimal Scaling appartenente alla famiglia PRINCALS; cfr. De Leeuw & Van Rijkevorsel, 1980)¹.

3.1 Principi metodologici della procedura CATPCA

In un algoritmo ALSOS ogni variabile osservata viene quantificata come prodotto di un vettore $\omega_j = [\omega_{j1} \dots \omega_{jh} \dots \omega_{jk_j}]'$ di parametri di *scaling* (le cui stime $\hat{\omega}_{jh}$ originano le variabili quantitative ricercate) e di una matrice-indicatore² G_j , nota perché tratta dai valori osservati. CATPCA utilizza, come metodo di quantificazione delle m variabili categoriche, proprio le componenti principali delle variabili trasformate, ottimizzate nello spazio fattoriale p -dimensionale, con $p \leq m$ (cfr. De Leeuw & Meulman, 1986; Meulman & Heiser, 1999). La procedura di ottimizzazione, nel caso più semplice (nessuna ponderazione per casi o variabili e nessuna variabile supplementare), stima in modo iterativo i parametri di *scaling* ω_j minimizzando la funzione

$$\sigma(\mathbf{Y}; \mathbf{\Omega}) = n^{-1} \sum_j \text{tr}[(\mathbf{Y} - \mathbf{G}_j \mathbf{\Omega}_j)' \mathbf{M}_j (\mathbf{Y} - \mathbf{G}_j \mathbf{\Omega}_j)],$$

ove la matrice \mathbf{M}_j è una matrice diagonale (di rango $n \times n$) i cui elementi diagonali valgono 0 se la j -ma osservazione è mancante o appartiene a categorie non valide, 1 in caso contrario; la matrice \mathbf{Y} (d'ordine $n \times p$) rappresenta le variabili riscalate in p dimensioni (punteggi fattoriali), normalizzate e centrate: ponendo $\mathbf{M} = \sum_j \mathbf{M}_j$ e \mathbf{u} = vettore-unità di dimensione n ,

$$\mathbf{Y}' \mathbf{M} \mathbf{Y} = n \mathbf{I}_p, \quad \mathbf{u}' \mathbf{M} \mathbf{Y} = \mathbf{O}.$$

L'algoritmo parte da una stima iniziale di \mathbf{Y} che soddisfi tali vincoli (salvo diversamente specificato, numeri casuali standardizzati e centrati); i vettori \mathbf{a}_j dei pesi di componente sono inizialmente calcolati come prodotto incrociato fra $\hat{\mathbf{Y}}$ e le variabili originali (codifiche categoriali) centrate e riscalate: $\mathbf{x}_j = [\mathbf{I}_n - \mathbf{M}_j \mathbf{u} \mathbf{u}' / (\mathbf{u}' \mathbf{M}_j \mathbf{u})] \mathbf{x}_j$, con $j = 1, 2, \dots, m$.

¹ Purtroppo, i prerequisiti di questa tecnica robusta e flessibile (tra cui, in primo luogo, la verifica che l'intervento dei valori mancanti sull'insieme delle variabili osservate non riduca eccessivamente il numero di casi validi, ossia quelli su cui sono disponibili valori non nulli di tutte le variabili) rendono improponibile la sua applicazione a dataset in cui molte variabili presentano dati mancanti: per esempio, nel pur numeroso dataset della *European Social Survey 2017*, i casi validi per l'insieme di *tutte* le 215 variabili attinenti il concetto di Smart Cities sono esattamente zero (rimozione inrocata "listwise"). Utile, dunque, la rimozione delle variabili maggiormente carenti da questo punto di vista.

² Una matrice-indicatore G_j (di dimensioni $n \times k_j$) è data, per ogni j -esima variabile osservata ordinale o nominale con k_j modalità, dai vettori colonna \mathbf{g}_{jh} ($h=1, 2, k_j$) definiti a loro volta, per ogni i -esima unità rilevata ($i=1, 2, n$), da uno scalare g_{ijh} che assume il valore 1 oppure 0 a seconda che detta unità rientri o non rientri nella h -esima categoria della variabile stessa. Si noti che le variabili scritte in carattere grassetto sono vettori e matrici. A causa degli assunti su cui si basano i metodi di stima (che richiedono appunto che la matrice dei dati sia espressa in punteggi interi, identificatori delle citate categorie), le eventuali variabili misurate su scala di rapporti o di intervalli vengono preliminarmente *discretizzate*. È anche indispensabile provvedere al trattamento delle osservazioni mancanti, preferibilmente con rimozione "listwise".

Il primo passo dell'iterazione consiste nel calcolare, data $\mathbf{D}_j = \text{diag}(\mathbf{G}'_j \mathbf{G}_j)$,

$$\hat{\mathbf{\Omega}}_j = \mathbf{D}_j^{-1} \mathbf{G}'_j \hat{\mathbf{Y}}. \quad [1]$$

Si ottengono così le prime quantificazioni di categoria, che in caso di variabili nominali sono calcolate direttamente tramite i pesi di componente: $\hat{\omega}_j = \hat{\mathbf{\Omega}}_j \mathbf{a}_j$. Se, invece, le variabili originarie sono ordinali (o numeriche, obbligatoriamente discretizzate), le $\hat{\omega}_j$ sono ottenute tramite una regressione monotona delle $\hat{\mathbf{\Omega}}_j \mathbf{a}_j$, ponderata con gli elementi diagonali di \mathbf{D}_j , che vengono poi ortogonalizzate con

$$\hat{\omega}_j^\perp = \hat{\omega}_j \sqrt{n / (\hat{\omega}'_j \mathbf{D}_j \hat{\omega}_j)}$$

allo scopo di calcolare i nuovi pesi di componente: $\mathbf{a}_j = (\mathbf{\Omega}'_j \mathbf{D}_j \hat{\omega}_j^\perp) / n$.

Tramite la matrice normalizzata delle stime di scaling $\hat{\mathbf{\Omega}}_j^\perp = \hat{\omega}_j^\perp \mathbf{a}'_j$ si può ora calcolare la matrice $\hat{\mathbf{Y}} = [\mathbf{I}_n - \mathbf{M} \mathbf{u} \mathbf{u}' / (\mathbf{u}' \mathbf{M} \mathbf{u})] (\sum \mathbf{M}_j \mathbf{G}_j \hat{\mathbf{\Omega}}_j^\perp)$. Dopo la verifica della convergenza della funzione obiettivo al minimo, si procede a chiudere il ciclo oppure, dopo aver ortogonalizzato $\hat{\mathbf{Y}}$ tramite *Singular Value Decomposition*, a riprendere l'algoritmo ripartendo dalla [1].

Le stime $\hat{\omega}_j$ relative all'ultima iterazione costituiscono, finalmente, le quantificazioni cercate delle varie categorie, tramite cui calcolare le ${}^{\text{os}}\mathbf{x}$. Dette variabili "ottimamente scalate" sono considerabili determinazioni di variabili quantitative continue.

3.2 Utilizzo e risultati della procedura CATPCA con eliminazione backward-stepwise

Avviata la procedura CATPCA sulla base delle 48 variabili residue delle precedenti esplorazioni uni e bivariate, a cui si va ad aggiungere la variabile di interesse, la sequenza di operazioni effettuate è la seguente:

- 1) le componenti principali dell'insieme delle variabili identificate sono selezionate tramite il criterio di estrazione più comune, ossia selezionando le componenti a cui corrispondono autovalori maggiori di 1, e dunque maggiormente rappresentative delle variabili osservate³;
- 2) valutando le comunanze (ossia le percentuali di varianza spiegata dalle componenti estratte) delle singole variabili osservate, viene identificata la variabile con la minor comunanza, che risulta dunque quella meno *determinata* dalla matrice di correlazione multipla, e detta comunanza viene confrontata con un valore soglia (generalmente fissato al 50%, cioè 0,5, ma che in questo caso sarà dato dal valore assunto volta per volta dalla variabile di interesse "country");
- 3) se la variabile identificata al passo precedente presenta un valore di comunanza inferiore alla soglia prefissata, viene rimossa dalla lista, altrimenti la procedura termina;
- 4) viene reinserita l'ultima delle variabili rimosse in precedenza, per estrarre le componenti principali della lista aggiornata e valutarne le comunanze, identificando quella con la comunanza minore: se la variabile identificata è differente da quella appena reinserita, si rientra nella procedura decisionale al passo 3;
- 5) altrimenti si valuta se il valore di comunanza di detta variabile è inferiore al valore soglia, il che implica la sua nuova rimozione dalla lista, per poi tornare al passo 1; in caso contrario, la procedura termina.

Le variabili superstiti nella lista al termine della procedura (dopo uno, due o decine di cicli) sono quelle che appartengono al medesimo sistema di correlazione multipla identificato

³ Gli autovalori λ_i costituiscono le radici dell'equazione caratteristica $|\mathbf{R} - \lambda \mathbf{I}| = 0$, ove \mathbf{R} è la matrice quadrata delle mutue correlazioni delle variabili osservate e \mathbf{I} la matrice identità di pari rango; essi corrispondono alla varianza delle rispettive componenti ma anche al contributo di tali componenti alla varianza complessiva di tutte le variabili osservate standardizzate. Poiché la varianza da ogni variabile standardizzata è uguale a 1, scegliere le componenti con autovalori maggiori di 1 corrisponde quindi a scegliere quelle che spiegano una varianza maggiore di quella introdotta nel modello da ogni singola variabile.

dall'Analisi delle Componenti Principali Categorie, di cui fa parte logicamente anche la variabile di localizzazione *Country* su cui saranno sviluppate tutte le successive osservazioni, analisi e previsioni. Esse garantiscono, peraltro, una maggiore robustezza dei risultati in quanto meno affette dai vari problemi connessi ai dati mancanti.

Nel caso presente, si tratta delle seguenti 36 variabili (oltre a *Country*):

1. Allow many/few immigrants from poorer countries outside Europe;
2. Allow many/few immigrants of different race/ethnic group from majority;
3. Allow many/few immigrants of same race/ethnic group as majority;
4. Belonging to particular religion or denomination;
5. Contracted hours per week in job-overtime excluded;
6. Contracted hours per week in job-overtime included;
7. Domicile description;
8. Employment contract unlimited or limited duration;
9. Employment relation;
10. Every 100 people in working age, how many are unemployed and looking for work;
11. Father born in [country];
12. Father's highest level of education (ES – ISCED);
13. Feeling about household's income nowadays;
14. Feeling safety walking alone in local area in dark;
15. How likely not enough money for household necessities next 12 months;
16. How much electricity in [country] should be generated from coal;
17. How much electricity in [country] should be generated from natural gas;
18. How much electricity in [country] should be generated from solar power;
19. How much electricity in [country] should be generated from wind power;
20. How often pray;
21. How worried, [country] too dependent on energy imports;
22. How worried, energy too expensive for many people;
23. Important that government is strong and ensures safety;
24. Important to be humble and modest, not draw attention;
25. Important to be successful and that people recognize achievements;
26. Important to get respect from others;
27. Important to have a good time;
28. Important to live in secure and safe surroundings;
29. Important to seek fun and things that give pleasure;
30. Important to show abilities and be admired;
31. Important to understand different people;
32. Legal marital status;
33. Mother born in [country];
34. Mother's highest level of education (ES – ISCED);
35. Political system allows people to have a say in what government does;
36. Political system allows people to have influence on politics.

Logicamente, nel questionario ESS, al termine parentetico *[country]* è sostituito il nome del Paese ove risiedono gli intervistati.

4. Breve analisi descrittiva delle variabili più correlate al Paese di residenza

Nella Fig. 1 è riportata l'opinione dei rispondenti in merito all'ingresso nel proprio Paese di immigrati provenienti da Paesi poveri extra-europei: opinione complessivamente poco favorevole nella maggior parte dei Paesi osservati. In particolare, l'Ungheria (HU), la Svizzera (CZ)

e la Federazione Russa (RU) si rivelano i Paesi più restii, rispettivamente con il 94,2%, il 76,8% e il 73,4% di cittadini favorevoli all'ingresso di pochi immigrati (*a few*) o di nessuno (*none*). Di contro, in Svezia (SE), Israele (IS) e Norvegia (NO) si riscontrano le percentuali più elevate, rispettivamente, con l'86,9%, l'81,4% e il 75,5%, di rispondenti favorevoli all'ingresso di molti (*many*) o alcuni (*some*) immigrati.

Figura 1. Risposte al quesito “*Allow immigrants from poorer countries outside Europe?*”, ripartite per Paese dei partecipanti alla Survey.

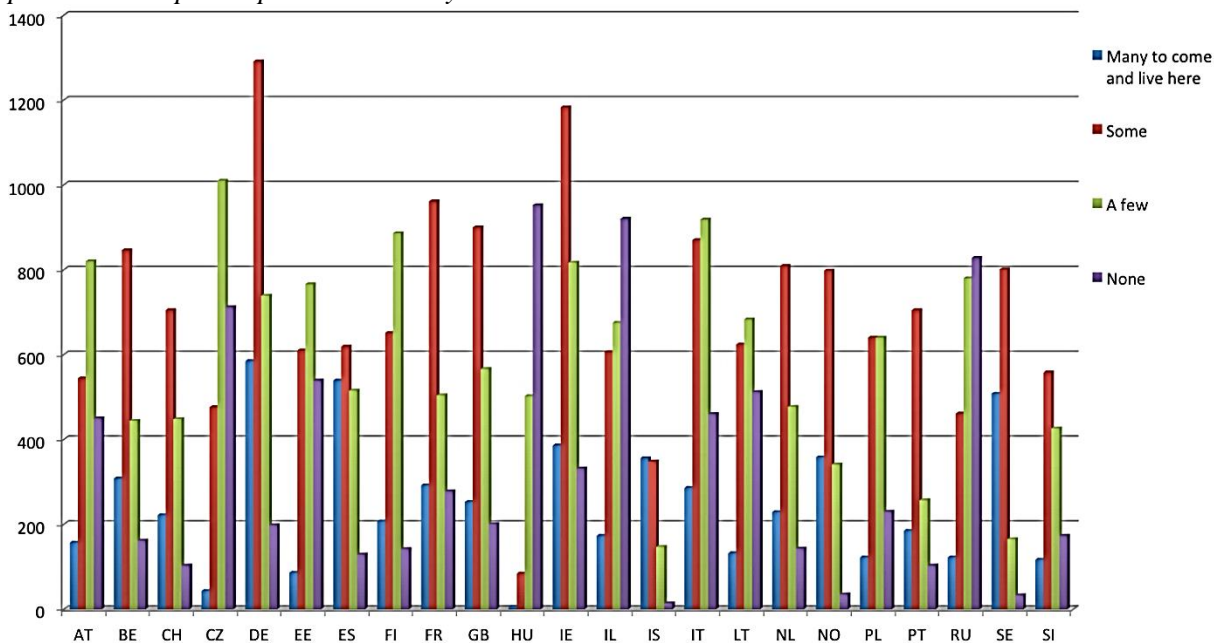
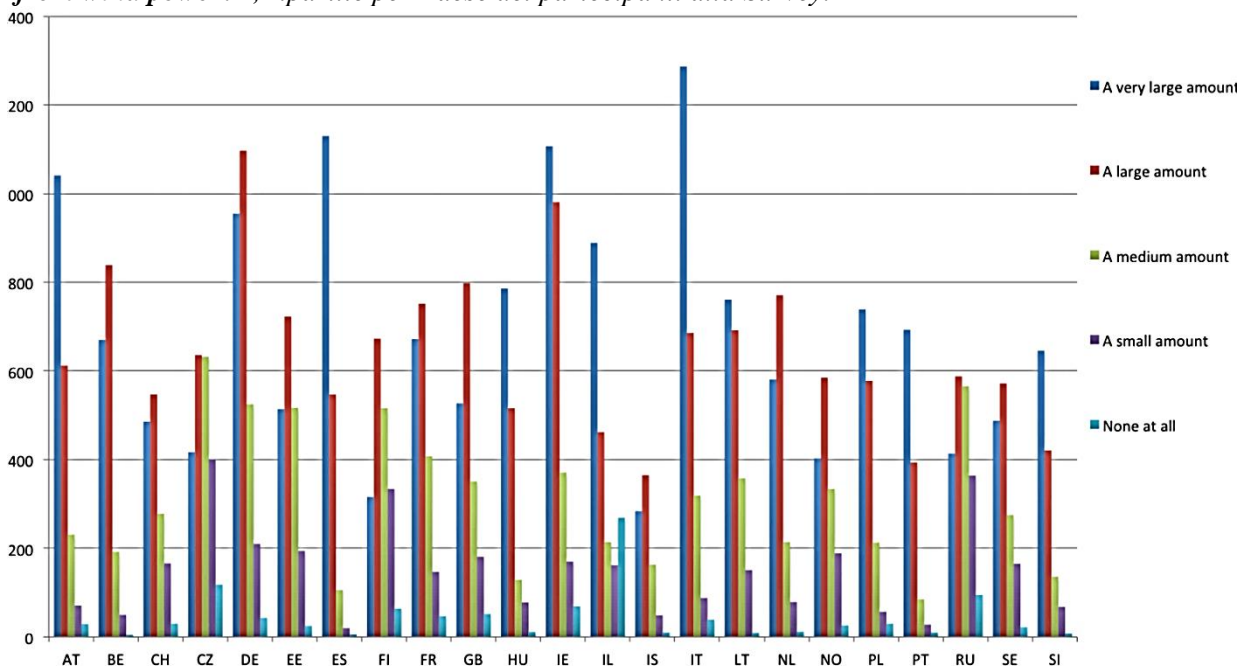


Figura 2. Risposte al quesito “*How much of the electric power used in [country] should be generated from wind power?*”, ripartite per Paese dei partecipanti alla Survey.



Per quanto riguarda la propensione all'utilizzo di fonti energetiche alternative, nello specifico, in tutti i Paesi si rileva un grande favore per l'utilizzo di energia eolica (Fig. 2). In particolare, Spagna (ES) e Portogallo (PT) riportano le percentuali maggiori di utilizzo di energia

eolica, rispettivamente, con il 92,7% e l'89,8% di rispondenti che affermano che il proprio Paese dovrebbe utilizzare enormi (*very large*) o grandi (*large*) quantità di energia eolica. In Svizzera (CZ) e Federazione Russa (RU), invece, non pochi rispondenti affermano che il proprio Paese dovrebbe generare energia eolica in misura scarsa o nulla (*small / none*), rispettivamente il 23,4% e il 21,9%.

Differenze altrettanto interessanti ed eloquenti si rilevano per tutte le variabili selezionate, ma una disamina puntuale delle distribuzioni bivariate esulerebbe dai limiti e dagli scopi del presente articolo.

Invece, risulta di cospicuo interesse valutare quanto l'intero insieme delle variabili qui identificate sia correlato alla residenza degli intervistati, ossia se e quanto il fatto che un soggetto risieda in un Paese invece che in un altro sia connesso all'espressione di determinate opinioni, al punto che conoscendo dette opinioni si possa ipotizzare, con un certo margine di errore, ove risieda il soggetto medesimo. In altri termini, se alla generalmente presumibile influenza del Paese di residenza e delle sue caratteristiche e usanze sulle opinioni dei suoi cittadini corrisponda una relazione statistica espressa in senso contrario: come se le opinioni espresse determinino (in realtà, rivelino) il Paese ove l'intervistato risiede.

Per far ciò, non potendo supporre che siano tutte dello stesso tipo le relazioni intercorrenti tra le caratteristiche e opinioni rilevate e la localizzazione territoriale degli intervistati, si è fatto ricorso a una tecnica esplorativa indipendente dalla forma delle relazioni medesime ed anche dalla gerarchia delle stesse: lo studio delle Reti Neurali Artificiali, per i cui riferimenti metodologici cfr., ad es., Orchard & Phillips 1991, Haykin 1998.

5. Valutazione delle variabili selezionate tramite Rete Neurale

L'utilizzo delle Reti Neurali ha visto un significativo miglioramento nelle performance a seguito della disponibilità di dataset su larga scala e del miglioramento delle capacità computazionali (cfr. Goodfellow *et al.*, 2016). Gli elementi fondamentali che costituiscono una Rete Neurale sono gli input, i valori in output e gli strati nascosti che consentono di trasformare gli input in output (cfr. Pham *et al.*, 2017).

L'obiettivo principale delle Reti Neurali è trasformare i dati in input in nuovi fattori che diventano linearmente separabili in classi (cfr. Roy *et al.*, 2019). Tuttavia, la maggior parte dei problemi che le Reti Neurali cercano di risolvere sono di tipo non-lineare e piuttosto complessi nella realtà. A tal fine, le Reti Neurali si basano sull'utilizzo di particolari funzioni, definite funzioni di attivazione, che consentono di convertire l'input in output mediante una procedura di tipo non-lineare (cfr. Agostinelli *et al.*, 2014). Senza le funzioni di attivazione, le Reti Neurali sarebbero delle semplici funzioni lineari. Pertanto, queste costituiscono la colonna portante di ogni architettura basata su Reti Neurali.

In questo studio sono state provate e confrontate vari algoritmi e architetture di Reti Neurali Artificiali nella ricerca di una struttura predittiva stabile ed efficace, definendo come variabile di output (come accennato) il Paese di residenza e come variabili di input tutte quelle prima identificate tramite analisi CATPCA. Sulla base delle sperimentazioni effettuate, l'architettura più adeguata risulta una *rete di Kohonen*, nota anche come *self-organizing map* (SOM).

La SOM è un'architettura di apprendimento non supervisionato in grado di produrre una rappresentazione discretizzata dello strato di input del campione di *training*. Questa architettura si differenzia dagli altri tipi di Reti Neurali in quanto si basa su una tecnica di apprendimento competitivo (cfr. Wehrens & Buydens, 2007) invece che sulla semplice correzione dell'errore mediante retro-propagazione e discesa del gradiente. La SOM è particolarmente adatta alla rappresentazione di dati in input di dimensione elevata.

Basandosi sul calcolo della distanza euclidea, viene esaminato ogni neurone per individuare quello che presenta il vettore con maggiore similarità rispetto al vettore in input. Il neurone vincente è noto come *Best Matching Unit* (BMU). L'architettura SOM prevede, inoltre, che anche i neuroni vicini abbiano vettori simili. Il neurone BMU e i neuroni vicini sono aggiornati ad ogni iterazione al fine di massimizzare la similarità rispetto al vettore d'ingresso. L'aggiornamento del vettore dei pesi $W_i(t)$ del neurone i alla t -esima iterazione avviene sulla base del differenziale

$$DW_{j,i} = \alpha(t)q_{j,v}(t)(x_i - w_{j,i}),$$

dove v rappresenta l'indice del neurone vincente BMU per il vettore d'ingresso x_i , $\alpha(t)$ rappresenta un coefficiente di apprendimento che decresce nel tempo, $q_{j,v}(t)$ rappresenta la funzione di vicinato che restituisce la distanza tra il neurone v e il neurone j alla t -esima iterazione. L'algoritmo termina dopo un numero predeterminato di iterazioni.

L'architettura SOM può essere utilizzata anche per apprendimento supervisionato, quindi è possibile inserire nel modello delle variabili addizionali che possono essere trattate come variabili dipendenti per le quali ottenere una previsione. Appunto questa opzione viene qui utilizzata per rappresentare i Paesi di residenza dei partecipanti all'indagine ESS 2017.

La procedura è stata implementata su un campione di training pari all'80% dei casi validi (circa 21.000 soggetti). Tale rete, nella *training set*, riesce a identificare correttamente i Paesi di residenza degli intervistati nel 98,6% dei casi, ed altrettanto avviene nella *test set*: risultato indubbiamente eccezionale, soprattutto considerando che una attribuzione casuale di una delle 22 residenze ai soggetti partecipanti risulterebbe corretta in meno del 5% dei casi.

In effetti, il contributo delle variabili identificate alla corretta identificazione del Paese di residenza degli intervistati (basato sul loro livello di correlazione con l'output) risulta distribuito in modo abbastanza graduale ma omogeneo.

Le variabili maggiormente rilevanti nell'identificazione della residenza risultano: “*Every 100 working aged, how many unemployed and looking for work*”, “*How much electricity in [country] should be generated from coal*”, “*How worried energy is too much expensive for many people*”, “*How worried, [country] is too much dependent on energy imports*”.

Invece, le variabili meno rilevanti sono: “*Total contracted hours per week in job-overtime excluded*”, “*Total contracted hours per week in job-overtime included*”, “*Legal marital status*”, “*Employment relation*”.

Volendo incrementare il potere discriminante della Mappa di Kohonen e contestualmente la parsimonia della rete, risulta opportuno rimuovere via via (con un procedimento di *backward elimination*) le variabili meno determinanti, fino ad ottenere, per una data combinazione di variabili *input*, il massimo livello di importanza (o determinazione, derivata dai coefficienti di correlazione input-output) e il minimo errore di classificazione dell'output. Tale condizione ottimale si verifica rimuovendo dapprima “*Employment relation*”, poi “*Legal marital status*” e infine entrambe le variabili occupazionali citate. Nessuna delle restanti variabili risulta irrilevante ai fini dell'identificazione territoriale (cfr. Tab. 2).

La rete così adattata classifica in maniera errata solo 10 soggetti su 16.644 campionati nella *training set* (errore: 0,0006) e 2 intervistati su 4.159 del *test set* (errore: 0,0005).

Per valutare la robustezza della configurazione ottenuta, infine, è stata condotta una procedura di *cross-validation* con *10-fold*, che ha pienamente confermato i precedenti risultati: previsione corretta nel 99,992% dei casi appartenenti alla *training set* e nel 99,991% dei casi nella *test set*.

È lecito supporre, quindi, che la quasi perfetta relazione (logicamente non lineare) tra opinioni espresse e Paesi di residenza possa consentire, in una successiva ricerca, di ricostruire in modo preciso e considerare rappresentative e generalizzabili le relazioni intercorrenti tra opinioni e indicatori oggettivi rilevati per i medesimi Paesi in tema di Smart Cities.

Tabella 2. *Pseudo-Importanza* delle variabili di opinione dell'indagine European Social Survey (ESS) 2017 nell'identificazione dei Paesi di residenza degli intervistati tramite rete di Kohonen.*

Variabili	Impor- tanza	Importanza normalizzata
Every 100 working aged, how many unemployed and looking for work	0,058	100,0%
How much electricity in [country] should be generated from coal	0,048	83,4%
Allow many/few immigrants of different race/ethnic group from majority	0,047	80,9%
Allow many/few immigrants from poorer countries outside Europe	0,047	82,3%
How worried, energy too expensive for many people	0,047	82,3%
Feeling about household's income nowadays	0,047	80,9%
How worried, [country] too dependent on energy imports	0,046	79,1%
Political system allows people to have influence on politics	0,041	71,1%
How much electricity in [country] should be generated from natural gas	0,037	64,5%
Political system allows people to have a say in what government does	0,036	63,1%
Important to get respect from others	0,035	60,0%
How likely not enough money for household necessities next 12 months	0,034	58,6%
Important to live in secure and safe surroundings	0,033	57,3%
Allow many/few immigrants of same race/ethnic group as majority	0,030	51,5%
Belonging to particular religion or denomination	0,030	52,9%
Important to be humble and modest, not draw attention	0,030	51,5%
Important to be successful and that people recognize achievements	0,030	51,6%
How often pray	0,029	49,7%
Important that government is strong and ensures safety	0,029	50,5%
Feeling safety walking alone in local area in dark	0,028	48,0%
How much electricity in [country] should be generated from solar power	0,026	44,5%
Domicile, respondent's description	0,025	43,5%
How much electricity in [country] should be generated from wind power	0,024	41,1%
Important to have a good time	0,023	39,8%
Mother's highest level of education, ES - ISCED	0,022	38,8%
Important to show abilities and be admired	0,022	39,0%
Father's highest level of education, ES - ISCED	0,021	37,0%
Important to understand different people	0,021	36,4%
Important to seek fun and things that give pleasure	0,020	34,2%
Employment contract unlimited or limited duration	0,015	26,5%
Father born in country	0,010	18,2%
Mother born in country	0,010	18,0%

* Calcolata sulla base del coefficiente di correlazione non lineare tra le variabili di input e di output.

6. Conclusioni

L'articolo qui presentato, limitato per motivi editoriali alla disamina dei concetti fondanti della ricerca in corso e alla descrizione dei criteri metodologici di scelta delle variabili di interesse, espone i primi riscontri di una approfondita analisi dei fondamenti sociali, economici e politici relativi alla diffusione ed evoluzione del concetto di "smart city" in Europa.

Il successo delle iniziative in ambito "smart city" dipende, tra gli altri fattori, dalla capacità dei governi di tener conto delle caratteristiche specifiche del contesto ambientale e sociopolitico in cui tali iniziative sono intraprese. A tal fine, nel presente lavoro l'attenzione è stata focalizzata sull'analisi dei dati raccolti attraverso la European Social Survey (ESS) 2017 che riguarda le opinioni dei cittadini residenti in 22 Paesi, per la maggior parte compresi nell'Unione Europea, su tematiche di carattere economico, sociale, demografico, politico, ambientale.

Delle 534 variabili di partenza, sono state prese in considerazione le variabili strettamente connesse con le sei dimensioni riconosciute in ambito “smart city” a livello accademico (*people, governance, economy, mobility, environment e living*). Attraverso successive analisi è stato estrapolato un sottoinsieme di 36 variabili particolarmente significative in relazione al contesto territoriale di appartenenza. La rilevanza di queste variabili in relazione al territorio è stata verificata tramite una “inversione” della relazione stessa, ossia supponendo che le opinioni espresse dagli intervistati (nella prima analisi considerate *dipendenti* dal Paese di residenza) siano in grado *da sole* di identificare in misura abbastanza precisa il territorio a cui afferisce il singolo rispondente. I risultati ottenuti da questa “validazione inversa” rivelano che ciò avviene nella quasi totalità dei casi.

Il presente lavoro si configura in ogni caso come un’analisi preliminare e indispensabile ai fini di una successiva ricerca la quale, prendendo in considerazione le 32 variabili risultate discriminanti a livello territoriale, le porrà in relazione ad un set di indicatori quantitativi prodotti da Eurostat in ambito *Digital Economy and Society* per ciascun Paese considerato. Lo scopo finale dell’analisi è, infatti, quello di comprendere la relazione tra la percezione (assolutamente soggettiva) dei cittadini su tematiche “smart city” e la misura oggettiva del livello di “smartness” a livello territoriale.

Riferimenti bibliografici

- Agostinelli, F.; Hoffman, M.; Sadowski, P; Baldi, P. (2014). *Learning activation functions to improve deep neural networks*. ICLR Workshop 2015: San Diego, CA, USA. In: arXiv preprint: <http://arxiv.org/abs/1412.6830>.
- Alawadhi, S.; Aldama-Nalda, A.; Chourabi, H.; Gil-Garcia, J. R.; Leung, S.; Mellouli, S.; Nam, T.; Pardo, T. A.; Scholl, H. J.; Walker, S. (2012). Building understanding of smart city initiatives. In: *International conference on electronic government*. Springer, Berlin, Heidelberg, pp. 40-53.
- Anthopoulos, L.; Janssen, M.; Weerakkody, V. (2019). A Unified Smart City Model (USCM) for smart city conceptualization and benchmarking. In: *Smart Cities and Smart Spaces: Concepts, Methodologies, Tools, and Applications*. IGI Global, pp. 247–264.
- Barletta, V. S.; Caivano, D.; Nannavecchia, A.; Scalera, M. (2019). A Spell Checking Web Service API for Smart City Communication Platforms. *Open Journal of Applied Sciences*, 9(12), 819.
- Barletta, V. S.; Caivano, D.; Dimauro, G.; Nannavecchia, A.; Scalera, M. (2020). Managing a Smart City Integrated Model through Smart Program Management. In: *Applied Sciences*, 10(2), 714.
- Eger, J. M. (2009). Smart growth, smart cities, and the crisis at the pump a worldwide phenomenon. In: *I-WAYS-The Journal of E-Government Policy and Regulation*, 32(1), 47-53.
- De Leeuw, J.; Meulman, J.J. (1986). Principal component analysis and restricted multidimensional scaling. In: W. Gaul, M. Schader (eds.), *Classification as a tool of research*, Amsterdam (NL): 83-96.
- De Leeuw, J.; Van Rijckevorsel, J. (1980). Homals and Princals - Some Generalizations of Components Analysis. In: E. Diday, Y. Escoufier, L. Lebart, J.P. Pages, Y. Schektman, R. Tomassone (eds.), *Data Analysis and Informatics*, Amsterdam (NL): 231-241.
- Goodfellow, I.; Bengio, Y.; Courville, A. (2016). *Deep learning*. MIT press.
- Gretzel, U.; Sigala, M.; Xiang, Z.; Koo, C. (2015). Smart tourism: foundations and developments. In: *Electronic Markets*, 25(3), 179–188.
- Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*, 2nd ed. New York: Macmillan College Publishing.
- Odendaal, N. (2003). Information and communication technology and local governance: Understanding the difference between cities in developed and emerging economies. In: *Computers, Environment and Urban Systems*, 27(6), 585-607.
- Orchard G.A., Phillips W.A. (1991). *Neural Computation. A Beginner’s guide*. London: Lawrence Erlbaum Associates.
- Meulman, J.J.; Heiser, W.J. (1999). *Categories 10.0*. SPSS Inc., Chicago.

- Nam, T.; Pardo, T. A. (2011). Conceptualizing smart city with dimensions of technology, people, and institutions. In: *Proceedings of the 12th annual international digital government research conference: digital government innovation in challenging times*, pp. 282-291. ACM.
- Pham, B. T.; Bui, D. T.; Prakash, I.; Dholakia, M. B. (2017). Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. In: *Catena*, 149, 52-63.
- Roy, S. K.; Manna, S.; Dubey, S. R.; Chaudhuri, B. B. (2019). LiSHT: Non-Parametric Linearly Scaled Hyperbolic Tangent Activation Function for Neural Networks. In: *arXiv preprint arXiv:1901.05894*.
- Soomro, S.; Miraz, M. H. ; Prasanth, A.; Abdullah, M. (2018). *Artificial intelligence enabled IoT: Traffic congestion reduction in smart cities*. Doi: 10.1049/cp.2018.1381.
- Wehrens, R.; Buydens, L. M. (2007). Self-and super-organizing maps in R: the Kohonen package. *Journal of Statistical Software*, 21(5), 1-19.