Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 10 aprile 2024

Ricerca bibliografica

Articoli teorici

Indice	***Titolo e DOI***	***Informazioni sulla pubblicazione***	***Casi particolari delle proprietà statistiche - geometriche***	***Considerazioni in confronto con le altre metriche***
1	https://doi.org/10.1007/978-3-319-07491-7_5 "*Una famiglia di dataset con parametri bidimensionali e la sua applicazione per confrontare i diversi indici del cluster validation".*	15th Mexican Conference, 21-24 giugno 2023, Mexico	In questo articolo si analizzano due casi differenti attraverso quattro algoritmi di clustering: Davies-Bouldin Index, Dunn Index, Cluster silhouette width index, gap statistic. In questo articolo vengono utilizzati due importanti algoritmi di cluster per comparare i vari cluster validation indices: il clustering gerarchico con la distanza euclidea e il clustering K-means. Il dataset è composto da 30 dati artificiali bidimensionali suddivisi in quattro gruppi, ognuno con caratteristiche specifiche. Vengono usati i dati bidimensionali perchè sono più facili da essere visualizzati. Tutte e quattro le metriche vengono utilizzate sia nel caso gerarchico che nel caso k-means. Quando i cluster sono ben separati i risultati sono migliori (ex: Three); quando i cluster sono più complessi i risultati sono peggiori (ex: Bean).	Nell'articolo viene mostrata una tabella con i risultati delle varie metriche e il numero di volte in cui ogni indice ha dato un numero corretto di cluster. Da qui si può notare che Davies-Bouldin e Cluster Silhouette Width sono gli indici con i risultati peggiori, mentre Dunn e Gap statistic sono quelli con risultati migliori.
2	*"Optimized K-Means Clustering Model based on Gap Statistic"*	International Journal of Advanced Computer Science and Applications (livello Q3)	La gap statistic viene applicata oltre che al k-means classico anche all'optimized k-means.	

Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 10 aprile 2024

3	*"Temporal gap statistic: A new internal index to validate time series clustering"*	Chaos, Solitons and Fractals (livello Q1)	Viene introdotta la temporal gap statistic.	
4	https://doi.org/10.3390/j2020016 *"Research on K-Value Selection Method of K-Means Clustering Algorithm"*	MDPI: Multidisciplinary Digital Publishing Institute (livello Q2).	Non viene modificata.	Confronto con altre metriche
5	https://doi.org/10.1016/j.chaos.2020.110326 *"A comparison of Gap statistic definitions with and with- out logarithm function"*	Chaos, Solitons and Fractals (livello Q1)	Viene tolto il logaritmo dalla formula	
6	https://doi.org/10.5430/air.v7n1p15 *"Estimating the number of clusters using diversity"*	Journal of Artificial Intelligence Research (livello Q1)	Non viene modificata	Confronto tra le metriche
7	https://doi.org/10.1111/j.1541-0420.2007.00784.x	Libro	Viene modificata	

Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 10 aprile 2024

	"Determining the Number of Clusters Using the Weighted Gap Statistic"			
8	https://doi.org/10.1263/jbb.105.273 *"Modified Fuzzy Gap Statistic for Estimating Preferable Number of Clusters in Fuzzy k-Means Clustering"*	JOURNAL OF BIOSCIENCE AND BIOENGINEERING (livello Q2)	Viene utilizzata la fuzzy gap statistic	
9	http://hdl.handle.net/10919/29957 *"Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion"*	Libro	Viene usata la weighted gap statistic	
10	https://doi.org/10.1007/978-3-030-45778-5_15 *"A New Approach to Determine the Optimal Number of Clusters Based on the Gap Statistic"*	Libro	Problema: la gap statistic di Tibshirani non può essere applicata quando i cluster non sono ben separati. Come fare quando abbiamo cluster sovrapposti? Soluzione: viene introdotta una nuova gap. Viene calcolata la decelerazione della gap. Quando si ha una minima sovrapposizione dei cluster, la velocità di incremento da Gap(k) a	

Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 10 aprile 2024

			Gap(k+1) è minore rispetto all'incremento da Gap(k-1) a Gap(k), rendendo massima la Dacc statistic in $k = K$. Quando invece, i cluster sono tanto sovrapposti neanche la Dacc funziona. In questo caso si incorpora 1-standard error e si calcola la DaccSE. Con questo metodo si prende il k massimo locale: se da k a k-1 ho un decremento maggiore che da k+1 a k, il k da prendere come numero di cluster migliore è k. Se da k a k-1 ho un decremento minore rispetto che da k+1 a k, il k da prendere come numero di cluster migliore è k+1	

TABELLA RIASSUNTIVA ARTICOLI TEORICI

CASES	SPECIAL CASES AND VARIANS OF STATISTICAL PROPERTIES - GEOMETRIC	CONSIDERATION TO BE CONFUSED WITH OTHER METRICS
10 cases	6 cases out of 10 equal to 60%, yes. 4 cases out of 10 equal to 40%, no.	3 cases out of 10 equal to 30%, yes. 7 cases out of 10 equal to 70%, no.

Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 10 aprile 2024

Articoli applicativi non biomedici

Indice	***Titolo e DOI***	***Informazioni sulla pubblicazione***	***Casi particolari delle proprietà statistiche - geometriche***	***Considerazioni in confronto con le altre metriche***
1	https://doi.org/10.1016/0375-6742(89)90076-9 *"Confronto tra grafici di probabilità e statistica del gap nella selezione delle soglie per i dati geochimici di esplorazione"*	Journal of Geochemical Exploration (livello Q2), aprile 1989.	Si confrontano la gap statistic e il probability plots per scegliere la soglia per i dati geochimici. Molte tecniche per selezionare le soglie mettono a confronto la distribuzione dei dati osservati e un modello di distribuzione teorica. Di solito il modello di distribuzione teorica è quella normale. Il grado di oggettività di queste tecniche può variare notevolmente. Per questa ragione viene proposto un sistema di classificazione per le tecniche di selezione delle soglie che comprende: categorie Experiental, model based subjective e model based objective. Quest'ultimo comprende la gap statistic e il probability plots. La gap statistic prevede un confronto della distribuzione normale con la distribuzione geochimica osservata dai dati. Entrambe le tecniche vengono	In questo articolo, quindi, è stata studiata la gap statistic comparandola alla probability plots. I risultati dimostrano che entrambe le tecniche funzionano bene con la maggior parte dei dati. I probability plots forniscono risultati leggermente migliori rispetto alla gap statistic. Infatti, quando si hanno delle distribuzioni distorte (a causa per esempio di deviazioni standard diverse) la gap statistic risulta avere una precisione molto scarsa e una variabilità nella classificazione dei campioni molto elevata.

			comparate usando dei dati generati casualmente dal dataset.	
2	https://doi.org/10.1016/j.gexplo.2016.01.002 "*An extended local gap statistic for identifying geochemical anomalies*"	Journal of Geochemical Exploration (livello Q2), maggio 2016.	I vari processi geochimici vengono prodotti da processi geologici. Questi ultimi possono essere rappresentati da distribuzioni statistiche di diverse forme e parametri, dunque la distribuzione dei dati geochimici è un mix di sottopopolazioni che rappresentano anomalie. Bisogna trovare una soglia di classificazione in modo tale da etichettare le varie anomalie: a tale scopo viene utilizzata la gap statistic locale. La gap statistic considera solo le proprietà statistiche della frequenza all'interno delle distribuzione dei dati, trascurando le variazioni spaziali dei dati geochimici, che potrebbero dare informazioni preziose. Per questo motivo viene utilizzata la tecnica del "sliding window" , la quale prende in considerazione anche i dati vicini spaziali, ampliando la finestra. In questo studio la gap statistic e la "sliding window" collaborano,	Tra i vari metodi di localizzazione, la tecnica del "sliding window" è semplice ed efficace. Utilizzando questa tecnica, le statistiche del vicinato che vengono calcolate possono quantificare le variazioni spaziali e definire ulteriori dettagli. La gap statistic locale è una di queste statistiche del vicinato. La gap statistic locale è stata utilizzata usando una serie di "window sizes" e il valore del test Student è stato applicato per misurare la correlazione spaziale tra le anomalie derivate da ogni dimensione e le occorrenze del minerale conosciuto. La dimensione ottimale della finestra per calcolare la gap statistic è determinato quando il valore del test Student è il più grande. La gap statistic locale si è dimostrata molto utile e appropriata.

			creando così la gap statistic locale. Quest'ultima viene utilizzata per identificare anomalie.	
3	https://doi.org/10.1093/gji/ggac326 *"Comparative analysis of the optimum cluster number determination algorithms in clustering GPS velocities"*	Geophysical Journal International (livello Q1), 17 agosto 2022.	Il Gps viene utilizzato per analizzare il movimento delle placche. I campi di velocità derivati dal Gps possono essere utilizzati come base per l'analisi del clustering per creare una definizione preliminare della geometria dei blocchi. Le velocità del Gps vengono clusterizzate da 2 cluster a 10 e vengono utilizzati cinque diversi datasets. Successivamente si applicano le metriche Davies-Bouldin, elbow method, Gap e silhouette. I dati utilizzati riguardano: Turchia (Number 1), Turkish national permanent gps network and Marmara region continuous network (Number 2), Anatolia centrale (Number 3). I datasets creati sono dunque: Number 1, Number 1&2, Number 1&3, Number 2&3 e Number 1&2&3.	- Number 1: tutte e quattro le metriche danno come risultato 5 cluster. - Number 1&2: la gap dà il valore più alto pari a 7. - Number 1&3: tutte le metriche danno 5 come risultato - Number 2&3: la gap e la Davis-Bouldin danno il valore più alto cioè 8. - Number 1&2&3: tutte le metriche danno 5 come risultato. Da notare che, mentre 1 e 3 coprono la Turchia in modo omogeneo, il 2 copre il centro dell'Anatolia. Il cambio di velocità delle placche va a modificare i valori delle metriche. Il numero ottimale dei cluster è 5 e l'elbow method è l'unica metrica che su cinque dataset dà quattro volte quel risultato: vuol dire che l'elbow method non viene modificato dal cambio di velocità. Dal grafico della Gap statistic si

				vede chiaramente che dal valore 1 a 5 la curva cresce, dopo il 5 la curva non ha un incremento significativo. Nel caso della Davies-Bouldin si ha come valore ottimale 5 ma anche 8, infatti dal grafico si può notare che i valori sono tra loro molto vicini: il valore ottimale è 5 perchè bisogna prendere come risultato della metrica quello più vicino a 0. I risultati più appropriati nel caso della Silhouette sono quelli compresi tra 5 e 8: il numero di cluster ottimale è pari a 5 perchè si ha il valore della silhouette più vicino a 1.
4	10.1016/j.jneumeth.2020.108651 *"Determining the number of states in dynamic functional connectivity using T cluster validity indexes"*	Journal of Neuroscience Methods (livello Q2), 25 febbraio 2020, Atlanta.	Il clustering analysis è usato nello studio della connettività funzionale dinamica, poiché i dati vengono clusterizzati in un set di stati dinamici. Vengono usati, inoltre, vari metodi per determinare la migliore partizione dei clusters esistenti. Gli indici attualmente impiegati non forniscono una risposta chiara su quale sia il numero ottimale di clusters e, inoltre, vi è una	Vengono testati gli indici di validità dei clusters in modo da trovare il numero degli stati di connettività funzionale dinamica. Per fare ciò si utilizzano simulazioni e dati derivati dalla risonanza magnetica funzionale. Dei 24 metodi utilizzati, Davies-Bouldin e Ray-Turi sono quelli più idonei per

			mancanza di test da parte dei suddetti nel caso di dati di connettività funzionale dinamica. In questo articolo vengono considerati 24 metodi per trovare il numero ottimale dei clusters.	trovare il numero dei clusters sia nelle simulazioni, che nei dati reali. Elbow-Criterion, Silhouette e Gap statistic sono metodi ampiamente diffusi negli studi di connettività funzionale dinamica.
5	10.1109/ICICCSP53532.2022.9862439 *"Approcci per trovare il numero ottimale di cluster utilizzando le tecniche K-means e cluster gerarchico"*	2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP) (21-23 luglio 2022, India) → **da scaricare**	In questo articolo vengono messe a confronto diverse metriche per analizzare le loro performance: connettività, indice di Dunn, indice dei Silhouette. Vengono utilizzate alcune metriche per trovare il numero ottimale di cluster: elbow method, gap statistic, silhouette e cluster gerarchico, k-means. Si ha a disposizione un dataset di 25 oggetti. Hanno applicato il k-means con k che va da 2 a 7.	La gap statistic calcola 3 come numero ottimale di cluster, l'elbow method dà come valore 4 e la silhouette 2. Il k-means e il cluster gerarchico producono simili risultati.
6	10.3390/s23042345 *"Sensor Clustering Using a K-Means Algorithm in Combination with Optimized Unmanned Aerial Vehicle Trajectory in Wireless Sensor Networks"*	MDPI: Multidisciplinary Digital Publishing Institute - Sensors (Journals), (livello Q2), 20 febbraio 2023	In questo articolo viene spiegato l'utilizzo dei WSN, ovvero sensori wireless all'interno di una rete, nel monitoraggio dei dati ambientali, quali: temperatura, precipitazioni, umidità. Questi, generalmente, sono low cost e low power, ma il problema principale sta nella trasmissione del segnale dai sensori ai centri di controllo. Per questo motivo vengono	Applicando la gap statistic e il K-means si è potuto notare che il numero ottimale dei sottocluster è pari a 4. Successivamente, si è potuto notare che l'UAV decodifica con successo gran parte dei messaggi provenienti dai sensori wireless. Lo scopo di questo studio è stato raggiunto grazie alle soluzioni proposte e

			introdotti gli UAV, i droni, che vengono mandati nell'area geografica per raccogliere i dati dal sensore. Per distribuire un WSN, è stato proposto l'utilizzo del clustering gerarchico. Nella tipologia gerarchica, i clusters contengono due tipi di nodi, i clusters members e i clusters heads. Questi ultimi ricevono i segnali dai clusters members e lo mandano agli altri clusters heads o alla base. Gli autori propongono come algoritmo quello del K-means per la selezione dei clusters heads, in modo tale che essi vengano scelti in base alla loro posizione nel cluster e al loro livello di batteria residua. In questo articolo vengono applicati il K-means e la gap statistic per ottenere un numero ottimale di sottocluster.	ai risultati verificati con la simulazione di monte carlo e le analisi teoretiche. Lo studio ha fornito molti benefici anche dall'utilizzo dell'algoritmo k-means per il clustering dei sensori wireless. Alcuni problemi rimangono irrisolti.
7	10.1080/02664763.2019.1675606 *"A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of Bangladesh"*	Journal of applied statistics (livello Q2), 29 settembre 2019, università di Dhaka, Bangladesh.	In questo articolo vengono utilizzati l'algoritmo agglomerato, la gap statistic con k-means e la gap statistic estesa con il fuzzy c-means per identificare l'omogeneità delle regioni in termini di precipitazioni. La gap statistic è usata per determinare il numero ottimale di clusters in caso di precipitazioni annuali durante i pre-monsoni, monsoni e	In questo studio sono state analizzate 30 stazioni di pluviometri del Bangladesh, in un periodo che va dal 1977 al 2012. I risultati finali che sono stati trovati rimangono tali se la gap statistic viene applicata usando il k-means e la gap statistic estesa viene applicata usando il fuzzy c-means. I tre algoritmi creano lo

			post-monsoni. Successivamente, vengono identificati gli oggetti dei clusters.	stesso numero di clusters ma non la stessa dimensione e la stessa omogeneità all'interno del cluster stesso. In più, nessun cluster ottenuto dal fuzzy c-means è eterogeneo, ma al contrario, è stato trovato un cluster eterogeneo nel caso di k-means e nell'algoritmo agglomerato. Da questo si evince che il fuzzy c-means è l'algoritmo preferibile.
8	10.1002/ece3.9681 *"Can vegetation be discretely classified in species-poor environments? Testing plant community concepts for vegetation monitoring on sub-Antarctic Marion Island"*	Ecology and evolution (Journal) (livello Q1), 2023, Università sud-Africa.	In questo articolo si parla della classificazione dei vegetali, in particolare di quella sull'isola di Marion caratterizzata da una vegetazione povera di specie che subisce rapidi cambiamenti ambientali. Viene usata l'analisi di cluster per testare la capacità di classificare questa vegetazione povera di specie e metterla in relazione con le classificazioni precedenti. Vengono usati: silhouette, Dunn, connettività, Gap statistic applicati agli algoritmi: Ward gerarchico, divisive analysis, k-means non gerarchico e pam.	Come risultato finale si ottiene che l'algoritmo che performa meglio è Ward applicato con la Silhouette e l'indice di Dunn. Tuttavia, tutti i metodi di classificazione producono cluster altamente connessi e con una separazione debole.

Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 10 aprile 2024

TABELLA RIASSUNTIVA ARTICOLI APPLICATIVI NON BIOMEDICI

CASES	SPECIAL CASES OF STATISTICAL PROPERTIES - GEOMETRIC	CONSIDERATIONS TO BE CONFUSED WITH OTHER METRICS
7 cases of which: - 2 geochemicals equal to 28,6%. - 1 geophysical equal to 14,3% - 1 neuroscientific equal to 14,3% - 1 on wireless networks equal to 14,3% - 1 on the statistical analysis of precipitation equal to 14,3% - 1 on plants equal to 14,3%	1 case out of 7 equal to 14,3%	All articles equal to 100%

Articoli biomedici

INDICE	CASES	PATHOLOGY	BIOMEDICAL PROBLEM THE AUTHORS ARE TRYING TO SOLVE	DATA TYPE	HOW THE METRIC WAS USED, WITH WHAT MOTIVATION, AND WITH WHAT INTERPRETATION
1	https://onlinelibrary.wiley.com/doi/epdf/10.1080/13682820600806680 → "*Subtyping of children with developmental dyslexia via bootstrap*	Developmental dyslexia	The present study is a novel application of resampling (bootstrap aggregating or bagging) methods and the gap statistic to the subtyping of children with developmental dyslexia. Obiettivi ricerca:	tre datasets artificiali e una database clinico di test data standardizzati (8 test) da 93 bambini con dislessia dello sviluppo. Questa procedura viene	Nel caso clinico riportato, la gap statistic viene usata per trovare una misura di confidenza, utilizzata nel determinare il numero ottimale di cluster. Poichè nelle precedenti

Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 10 aprile 2024

	aggregated clustering and the gap statistic: comparison with the double-deficit hypothesis"		1: usare la gap statistic e il bagging con dati multivariati 2: confrontare i risultati ottenuti del cluster tramite la gap statistic con le ipotesi iniziali del doppio deficit.	ripetuta su 93 bambini senza disabilità nella lettura che vengono matchati per sesso e età.	analisi dei cluster non si riusciva ad ottenere una stabilità e una affidabilità dei risultati del cluster, hanno deciso di combinare la gap statistic con il bagging.
2	doi: 10.1007/s00125-021-05485-5 *"Comparison between data-driven clusters and models based on clinical features to predict outcomes in type 2 diabetes: nationwide observational study"*	Diabetes	The authors want to prove the existence of differing clusters within type 2 diabetes be validated and that the sub-stratification cluster is more useful than traditional methods to predict diabetes outcomes.	We used data from the Swedish National Diabetes Register and included 114,231 individuals with newly diagnosed type 2 diabetes	The gap statistic is used with k-means
3	doi: 10.1126/sciadv.abj0320 *"Classifying chronic pain using multidimensional pain-agnostic symptom assessments and clustering analysis"*	Chronic pain conditions	The authors suggest a reversal of the paradigm - instead of assessing patient - reported symptoms as features of the a - priori determined pain condition, they examined whether such symptoms may serve to classify current and predict future pain conditions. If confirmed, our approach could be used to support personalized and efficient treatment of individuals with	Data were collected using Stanford University's CHOIR, a registry- based, learning health care system that administers an electronic survey assessing self-reported demographic information, medical history, and multiple domains of health status in	They used gap statistic to determine the optimal number of clusters. An ideal solution will have a small within cluster sum of squares, and therefore a large gap statistic.

			chronic pain.	real-world clinical settings	
4	10.1111/aji.12818 *"Multivariate analysis of cytokine profiles in pregnancy complications"*	Pregnancy complications	The aim of this study is to use statistical approaches to study and quantify the connection between cytokine profiles and different categories of pregnancy complications as compared to normal controls.	Groups of women studied along with the number of patients (n), clinical history, and mean gestational age	Gap statistic with other statistical tools were used to compare cytokine data of normal vs anomalous groups of different pregnancy complications. The authors used the first PCA and complemented this method using gap statistic with k-means, which is a standard method for finding the number of subgroups in multivariate data and identifying subgroup membership for each sample.
5	10.3389/fmolb.2021.645024 *"Tumor Microenvironment Characteristics of Pancreatic Cancer to Determine Prognosis and Immune-Related Gene Signatures"*	Pancreatic cancer	In this study the authors developed a method to quantify the tumor microenvironment (an environment in which tumor cells are produced and inhabit).	In this study, a total of 177 patients with PC from The Cancer Genome Atlas (TCGA) cohort and 65 patients with PC from the GSE62452 cohort in Gene Expression Omnibus (GEO) were included	The TME was classified by k-means clustering and differentially expressed genes were determined. A combination of elbow method and the gap statistic was used to explore the likely number of distinct clusters.

Elisa Merigo email: e.merigo1@campus.unimib.it / merigo.elisa@gmail.com
Matricola: 869436 - 10 aprile 2024

6	10.1080/15592294.2018.1497387 *"Cumulative lifetime maternal stress and epigenome-wide placental DNA methylation in the PRISM cohort"*	Link between maternal psychosocial stress and the child's problems	In this study, the authors examined associations between maternal cumulative lifetime exposures to traumatic and non-traumatic stressors and epigenome-wide methylation in the placenta among women enrolled in a longitudinal ethnically diverse pregnancy cohort	Participants included women enrolled in the PRogramming of Intergenerational Stress Mechanisms (PRISM) study, a prospective preg- nancy cohort of mother-child pairs originally designed to examine how perinatal stress influences child development as well as examining the role of the placenta in environmental programming. In brief, 238 women were recruited from prenatal clinics during the first or second trimester.	The gap statistic is used to find the number of clusters.
7	10.1007/s11060-016-2174-1 *"Integrative analysis of diffusion-weighted MRI and genomic data to inform treatment of glioblastoma"*	Brain cancer, glioblastoma	In this study the authors used DW-MRI images. The main objective was to evaluate ADC as an imaging biomarker of the molecular subtype of GBM and its ability to stratify patients with GBM.	In this retrospective study, we analyzed the expression of 12,042 genes for 558 patients from The Cancer Genome Atlas (TCGA). Among these patients, 50 patients had magnetic resonance imag- ing (MRI) studies including diffusion	After clustered the resulting genes, the authors identified the number of cluster using the gap statistic in conjunction with k-means method by computing the difference between the within-cluster dispersion and the dispersion expected under a reference null distribution.

				weighted (DW) MRI in The Cancer Imaging Archive (TCIA).	
8	10.1016/j.visres.2011.12.006 *"A simple nonparametric method for classifying eye fixations"*	Eye fixations	The natural movement of an observer's gaze over a scene is discontinuous. The aim of this study was to exploit some of the more general distributional properties of eye movements to construct a simple, completely nonparametric method of classifying fixations.	The method was tested on data recorded from a video eye- tracker sampling at 250 frames a second while experimental observers viewed static natural scenes in over 30,000 one-second trials.	The method was primarily speed-based but the optimum speed thresholds for classifying saccades was derived automatically from the data for each observer and stimulus individually. The derivation was founded on gap statistic for identifying the optimum number of clusters in a set of data.
9	https://doi.org/10.1017/S0033291702006311 *"Searching for a Gulf War syndrome using cluster analysis"*	Gulf War syndrome	In this article the authors examined if the Gulf and non-Gulf veterans are distinguished by their patterns of symptom reporting. The aim of this study is to cluster together individuals who have similar clinicals profiles.	The study population consisted of three randomly selected military samples from the UK Armed Forces. A standardized self-report ques-tionnaire requesting sociodemograp hic, military and health information was sent to these individuals.	The gap statistic was applied in this study to suggest which number of groups or clusters best describe the data. The gap statistic suggested that the four, five and eleven cluster solution produced by the k-means procedure would be the most interesting.
10	https://doi.org/10.1007/s00125-024-06184-7 *"Identification*	Diabetes	The authors want to identify distinct gestational diabetes mellitus subtypes through cluster	In this cohort study, we analysed datasets from a total of 2682	The gap statistic was used to find the optimal number of clusters. When

	and validation of gestational diabetes subgroups by data-driven cluster analysis"		analysis using routine clinical variables and analyze treatment needs and pregnancy outcomes across these subgroups.	women with GDM treated at two central European hospitals.	using k-means, the gap method suggested k=1.

TABELLA RIASSUNTIVA ARTICOLI BIOMEDICI

CASES	PATHOLOGY	BIOMEDICAL PROBLEM THE AUTHORS ARE TRYING TO SOLVE	DATA TYPE	HOW THE METRIC WAS USED, WITH WHAT MOTIVATION, AND WITH WHAT INTERPRETATION
10 cases	- 1 case of developmental dyslexia, equal to 10%. - 1 case of Gulf war Syndrome, equal to 10%. - 1 case of eye fixation, equal to 10%. - 1 case of brain cancer (glioblastoma), equal to 10%. - 2 cases of pregnancy complications, equal to 20%. - 1 case of pancreatic cancer, equal to 10%. - 1 case of chronic pain, equal to 10%. - 2 cases of diabetes, equal to 20%.	In 40% of articles, the authors try to classify the problems, the patients or the pathologies. In cases of diabetes (20%), the authors want to predict and identify the panthology In cases of pregnancy complications (20%), the authors try to quantify the associations between maternal lifetime stress and child's problems and the connection between cytokine profiles and categories of complications. In 10% of the articles the authors want to evaluate the ACD and stratify patients. In 10% of the articles the authors want to quantify the tumor microenvironment.	- 1 case uses an artificial database and medical records, equal to 10%. - 6 cases use medical records, equal to 60%. - 1 case use MRI, DW-MRI and medical records, equal to 10%. - 1 case uses video, equal to 10%. - 1 case uses text, equal to 10%.	Motivation: In 40% of the articles gap statistic is the best for describe the data In 50% of the articles the motivation is not provided. The metric was used: 60% with k-means 10% with speed-based 10% with hierarchical clustering 10% with pam In the last 10% the use of the metric is not provided.