How to Compress Categorical Variables to Visualize Historical Dynamics

Fabio Celli[1,*]

[1]Research & Development, Maggioli SpA, via Bornaccino 101, 47822 Santarcangelo di Romagna

Abstract

This paper explores innovative Knowledge Discovery and Representation techniques for historical data within the field of Digital Humanities. In particular this research introduces Time-Resolved Variables, an information compression technique, to represent the evolution of categorical variables through time. This technique proves to be more effective than One-Hot Encoding with Principal Component Analysis in explaining the increase of social complexity from historical data. Moreover, this work highlights the potential of Time-Resolved Variables to enhance model explanation by means of correlation analysis and graph visualization. The result of this work, the Chronos dataset, is available online in a shared Spreadsheet for collaborative research and paves the way towards more transparent and trustworthy use of AI in history and cliodynamics.

Keywords

Knowledge Discovery and Representation, Data Compression, Cliodynamics, Model Explanation

1. Introduction and Related Work

Knowledge Discovery and Representation (KDR) is crucial for the scientific method in Digital Humanities [1] as it bridges the gap between unstructured data and theory, enabling data-driven hypothesis testing [2]. However, certain phenomena like missing historical records [3], particularly prevalent in historical data, introduce data integrity issues [4] due to the increasing scarcity of information as we look deeper into the past. Additionally, crowdsourcing historical data annotation is challenging [5] due to the subjective nature of historical interpretation. However, technological advancements have led to the development of KDR techniques to address challenges such as these. Specifically, there are three generations of Knowledge Discovery systems [6]. The first generation focused on collecting and querying data through large databases. The primary challenges in this phase were related to Knowledge Organization and Information Retrieval [7]. The second generation introduced powerful tools for extracting and visualizing patterns within data, enabling the reconstruction of events from unstructured sources and their presentation in the form of timelines and maps. For instance, it is possible to generate maps based on data about medieval trade routes [8] [9] or timelines illustrating the evolution of linguistic events derived from textual data [10]. The third generation, which has recently emerged, leverages Artificial Intelligence (AI) and Large Language Models (LLMs) to tackle data integrity issues. One application of this kind is the restoration of ancient inscriptions using AI [11]. However, a significant challenge remains: ensuring that AI models are transparent to humans. The ultimate goal is to create systems that are trustworthy, human-readable, computationally efficient, and capable of self-maintenance within a human-data-AI continuum [12].

Crucially, the field of cliodynamics is facing the same challenges and, among other research goals, addressed the problem of croudsourcing, producing Seshat, a valuable expert-compiled historical dataset that is suitable for computational analysis [13]. The basic concept of Seshat is to provide quantitative and semi-structured data about past societies, defined as political units (polities). It contains data from 35

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20-21 2025, Udine, Italy

*Corresponding author.

fabio.celli@maggioli.it (F. Celli)

https://github.com/facells/fabio-celli-publications (F. Celli)

0000-0002-7309-5886 (F. Celli)

sampling points equally distributed across the globe in a time window from roughly 10000 BC to 1900 CE and sampled with a time-step of 100 years. Seshat is designed for hypothesis testing. For example it has been used for comparing competing hypotheses about the evolution of social complexity, like theories that see social complexity as a product of organizational challenges due to environmental changes [14] or a product of strong normative beliefs in moralizing Gods [15]. A data-driven analysis on Seshat with dynamic regression models revealed a strong causal role played by a combination of increasing agricultural productivity and adoption of new military technologies [16]. The Seshat databank provides many dimensions that report the presence or absence of a cultural trait in a polity at a specific point in time, for example the presence/absence of smelting copper skills, fortifications, firearms, written literature, coins and many others. In Data Mining these fine-grained cultural dimensions can be treated computationally with One-Hot Encoding (OHE), a widely used approach for expressing the presence or absence of categorical features into numerical ones, represented as 1 or 0. However, OHE results in very sparse feature values [17] and this makes it difficult to extract patterns and generalize models from data for Knowledge Discovery. Moreover, OHE does not solve the problem of missing data. A common approach is to employ a compression technique like Principal Component Analysis (PCA) [18] to combine the one-hot encoded features into more general variables. However, this process converts the categorical variables into aggregated dimensions, very difficult to interpret. Among compression techniques for time series there are some based on dictionaries, some based on function approximation, some on sequential algorithms and more recent techniques based on autoencoders, but there is no compression technique for time series designed to be human-readable [19].

2. Scope of this work

This paper addresses the challenges of data integrity and knowledge representation in historical research by introducing a novel, human-readable information compression technique for time series: Time-Resolved Variables (TRVs). In essence, TRVs represent sequences of categories, ordered chronologically by their first known historical appearance, as numerical values corresponding to their position in the scale. For instance, a time-resolved variable for military technologies might assign the values 0.1 to stone/wood, 0.2 to copper, 0.3 to bronze, 0.4 to iron and so on. Interpolation can be incorporated using decimal numbers between these primary scale steps, signifying transitional periods between one level and the next. The values assigned to the steps in TRVs are numerical indexes that represent interpretable labels in a sequence. TRVs can encode two semantic dimensions: steps and duration between steps. Steps (i.e. 0.1 is stone, 0.2 is copper, 0.3 is bronze etc..) return an equidistributed sequence, that is indicated for visualization with timeline charts and comparison of different scales (i.e. cultural evolution of different societies). Instead encoding both the sequence and duration between steps (i.e. -10 is stone/wood, -0.5 is copper, -0.3 is bronze etc..) returns the inherent structure of historical events and, depending on the presence of patterns, cloud lead to a low-discrepancy sequence or to a purely random sequence, indicated for scatterplot visualizations. This paper aims to discover the characteristics of TRVs with equidistributed sequences in regression, correlation and visualization tasks, by comparing them against OHE-PCA. Experiments can provide evidence to understand the difference between the information extracted with the two techniques. It is possible to compare TRVs also against other compression techniques, but PCA has been selected because principal components are uncorrelated and reduce redundant information, which is optimal for the selected tasks, in particular for correlation analysis and data visualization. Further experiments with TRVs are left to future work.

TRVs can be created in two ways: manually, by using information from digital libraries, or automatically, by prompting LLMs to extract knowledge [20]. Both methods have drawbacks. Manual annotation is time-consuming, while LLM-generated TRVs can contain inaccuracies or "hallucinations". To effectively use TRVs in experiments and gain an advantage over OHE-PCA, it is possible to adopt a three-step approach: First, manually annotate a small dataset to create a reliable ground truth. Second, use LLMs to generate a larger dataset of TRV annotations. Finally, evaluate the accuracy of the LLM-generated data by comparing it to the ground truth using correlations or other statistical tests. In

general, TRVs offer a transparent approach to compressing historical data into a numerical format that is machine-readable and human-readable at the same time, while also enabling interpolation of missing data points. This goes in the direction of trustworthy and transparent human-data-AI continuum.

This paper focuses on the following research questions (RQs):

- RQ1: Are TRVs more informative than OHE-PCA features to compress historical information?
- RQ2: Are TRVs useful for visualizing and interpret the results of Knowledge Discovery?

To answer RQ1, this paper presents a comparison of the same independent variables, encoded with TRVs and OHE-PCA, to predict a set of dependent variables, and evaluates the results with the coefficient of determination R^2. To answer RQ2, this paper presents a comparison of correlation graphs extracted from the same features, but treated as OHE-PCA and TRVs, providing an explanation of the results. Prompt engineering for the generation of TRVs with LLMs is outside of the scope of this paper, as it is covered in Celli and Mingazov 2024 [20]. However, it is important to report here that Gemini 1.5 flash is capable of grounding its output with real references, thanks to the Google search engine [21]. This has a great potential for digital libraries, as it is possible to automatically link TRVs and existing text sources by means of LLMs.

The paper is structured as follows: in Section 3 there is a description of TRVs annotated on data and the experiments to answer to RQ1 and RQ2. Finally, Section 4 reports a discussion of the results, draws conclusions and outlines directions for future work.

3. Method and Experiments

A schema of the method adopted in this paper is depicted in Figure 1. First there is a data preparation process where Seshat is integrated and compressed into a structured dataset, processed with OHE and PCA, dubbed "Seshat-pca" (Section 3.1). Then there is the TRVs annotation phase, where the original categorical variables of Seshat are grouped and transformed into scales according to first time of appearance found in scholarly literature. This operation creates a new dataset, named "Chronos" (Section 3.2), that is compared against Seshat-pca to test the predictive power of TRVs and OHE-PCA compression methods and answer RQ1 (Section 3.3). Then there is a second comparison, with the extraction of a correlation matrix and visualization of correlation graphs form the two datasets (Section 3.4). Finally there is the analysis and discussion of the graph model to answer RQ2 (Section 4).

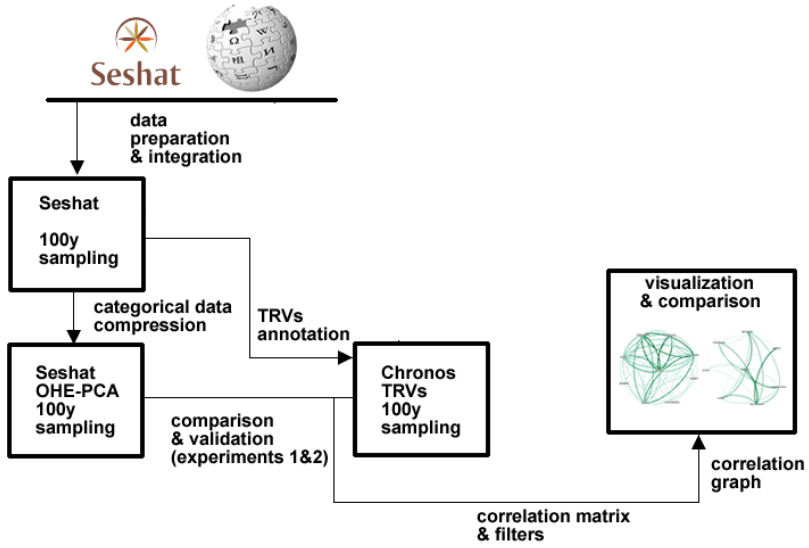

Figure 1: Schema of the method adopted.

The experiments are designed to test the usefulness of TRVs in the historical domain, but in principle this technique can be applied to any categorical data with a sequence, not necessarily temporal, and can preserve transparency while compressing information. Potential applications outside the historical domain include the analysis of music genres, where emotions can be expressed as scales [22], and hypothesis testing for theories that provide predictions which can be turned into sequences.

3.1. Compression of categorical variables in Seshat

The Seshat project released many replication datasets containing data about polities sampled from 35 Natural Geographic Areas (NGAs). The present work employed the categorical dimensions in the Social complexity dataset and the Axial age dataset [23]. The two datasets were structured and integrated, and the original variables were processed with OHE and PCA. The variables are grouped following the original macro-categories of Seshat:

military technology (militech): the PCA of presence/absence of copper, bronze, iron, steel, gunpowder siege artillery in a polity;

warfare tools and tactics (warfare): the PCA of presence/absence for each polity of small vessels, plate armor, laminar armor, settlement in defensive position, earth ramparts, moats, scaled armor, stone walls, horses, camels, spears, wood barks, leather cloth, shields, helmets, breastplates, limb protection, donkeys, composite bows, battle axes, daggers, swords, modern fortifications, bows, javelins, slings, crossbows, handled firearms, complex fortifications, fortified camps, chain mail tension siege engines, war clubs, elephants, ditch, pole arms, specialized military vessels, stone walls, atlatl;

agriculture technology (agritech): the PCA of presence/absence of cropping, field rotation, irrigation and fertilizers for each polity;

morality in religion (morality): the PCA of presence/absence of moral features in the polity: moral enforcement in this life, moralizing enforcement is agentic, moral religion is adopted by elites, moralizing is certain, broad moralizing norms, moral concern is primary, rulers are gods;

economy management (economy): the PCA of presence/absence of tokens, articles, precious metals, indigenous coins, foreign coins, paper currency;

information management (infomedia): the PCA of presence/absence of lists and tables, calendars, sacred text, religious literature, scientific literature, fiction, philosophy, practical literature, history;

writing system (alphabeth): the PCA of presence/absence of phonetic writing, non-phonetic writing, script, mnemonic devices, non-written records;

communication systems: the PCA of presence/absence of couriers, postal stations, general postal service in the polity;

infrastructure levels (infrastructure): the PCA of presence/absence of roads, mines and quarries, ports, canals, drinking water supply, irrigation and production systems, bridges, markets, food storage sites;

political system (politics): the PCA of presence/absence of constraints on executive by non-government (population representatives), constraints on executive by government (aristocracy), legal impeachment for each polity.

Finally, the three dependent variables of social complexity (that will be called also target variables) were processed in the same way as defined by Turchin[16]:

- *Social scale* (SCALE), is the PCA of the log-transformed polity population, polity territory and the population of the largest settlement.
- *Hierarchical complexity* (HIER), is the PCA of the raw count of levels in administrative, military and settlement hierarchies.
- *Specialization of governance* (GOV), is the PCA of the following 11 One-Hot encoded variables: presence or absence of professional officers, soldiers, priests, lawyers, full-time bureaucrats, specialized buildings for government, examination system, merit promotion, formal legal code, full-time judges, courts.

It is very important to point out that **Social scale** and **Hierarchical complexity** are originally homogeneous sets of numerical variables, while **Specialization of governance** is a heterogeneous set of One-Hot encoded categorical variables. All the features in Seshat Were normalized in order to align them with the target variables.

3.2. Annotation of Time-Resolved Variables in Chronos

To ensure that the categories represented by TRVs are compatible with the ones in Seshat, original categories were arranged in a chronological order according to their earliest historical appearance. Wikipedia and scientific literature have been taken as reference. A random subset of 186 polities from Seshat with information available for social complexity variables has been selected for the annotation. The TRV scales were defined as follows:

Technology for military purposes (militech) encodes the level of military technology of a polity through time and is related to the military use of metals and technologies. It encodes the following steps: 0.1 is *stone/wood/clay*, 0.2 is *copper*, 0.3 is *bronze*, 0.4 *iron*, 0.5 *steel*, 0.6 *gunpowder* and 0.7 *uranium/nuclear*.

Warfare tactics (warfare) encodes the level of military strategy of a polity. It encodes the following steps: 0.1 *hunters with projectiles*, 0.2 *armies and fortifications*, 0.3 *armies with animals and chariots*, 0.4 *armies with naval and siege forces*, 0.5 there are *armies with industrial forces* including machine guns, logistics and tanks, and then 0.6 is *cyberwarfare*.

Type of agriculture (agritech) encodes the stage of agriculture development through history. At level 0 there is *spontaneous* cropping, at 0.1 there is *swidden/slash-and-burn* agriculture, at 0.2 is *fallow* agriculture with irrigation, at 0.3 there are *two fields/crop rotation*, at 0.4 there are *nitrogen-fix/fertilizers* and at 0.5 there are *GMOs* (Genetically Modified Organisms).

Religious system (religion) puts along a scale the types of religions. Not single religions, but religion clusters. In Seshat this dimension is restricted to features about moralizing Gods but, since Savage[24] pointed out that complex societies precede moralizing Gods, this dimension has been changed, in order to test whether it captures social complexity better than morality. At the first level (0.1) we have the *cult of the dead and spirits*, 0.2 is the *cult of ancestors/family/totem*. 0.3 are the *fertility cults*, 0.4 are the *polytheisms*, 0.5 are *monotheisms*, 0.6 are *philosophies*, such as Buddhism, Confucianism and Taoism. At 0.7 there is *humanism*, that includes atheism, agnosticism and ideologies (like communism and capitalism).

Economic level (economy) encodes the economic advancements in history. At level 0 there are *subsistence and exogamy*, at 0.1 *tokens and barter*. at 0.2 *precious metals and weights*, at 0.3 there are *coins*, at 0.4 *paper currency* and at 0.5 there is the *stock market*.

Information management (infomedia) encodes the level of informative content that a culture needs to mediate and diffuse to keep its stability. At level 0 there are *mnemonic devices and oral tradition*, at 0.1 there is *symbolism*, at 0.2 there are *calendars and lists*, at 0.3 *religious, philosophical and scientific texts*; at 0.4 there is *fiction literature* and at 0.5 *news and opinions*.

Writing system (alphabeth) encodes the chronological order of appearance of alphabet types. At level 0 there is *no writing*, at 0.1 *script and logographic writing*, at 0.2 *syllabic and non-phonetic writing* and at 0.3 are *phonetic* alphabeths.

Communication systems (communication) encodes the evolution of long-distance communication. At level 0.1 there are *couriers*, at level 0.2 *networks of courier stations*, at 0.3 it appears a *centralized postal service*, at 0.4 *electric narrowcasting* like the telegraph, at 0.5 the *electric broadcasting* like radio and TVs, and at level 0.6 there is *computer mediated communication*.

Infrastructure Level (infrastructure) encodes the capability of a polity to extract, transport, and preserve goods. At level 0.1 there are *routes and quarries*, at 0.2 *storage sites and special buildings*. at 0.3 there are *irrigation and production systems*, at 0.4 there are *urban markets*, at 0.5 *portual systems* and at level 0.6 *logistics and telecommunications*. Level 0.7 is about *space stations*.

Political system (politics) encodes the development of the limits that are imposed to the rulers. At the base level there is a *sole ruler*, and no limits to the ruler's power. At level 0.1 decisions were taken in a collective *assembly*, at level 0.2 there are *representatives of the population*, that are needed in large

societies to put a constraint on the decisions taken by aristocratic assemblies and are the foundation of democracies. At 0.3 there is *legal impeachment*, that allows an authority to legally remove the powers from a ruler without physically killing him.

In the context of polities, TRVs represent generalized increasing stages of social complexity, and by design tend to reduce variety into broad classes. For example Christianity, Islam and Judaism are very different religions, but all fall under the umbrella of monotheism. The same for the beginning of symbolism: symbols produced in 9000 BC in Göbekli Tepe and symbols produced in Jiahu around 6000 BC are totally different things, but functionally they both represent the transition to a new level of complexity in information management. These TRVs are all considered independent variables in the experiments.

3.3. Experiments

The first experiment aims to predict the same dependent variables (SCALE, HIER, GOV) using the independent variables encoded with TRVs and OHE-PCA in the two different datasets. The polities were aligned, obtaining the same 186 polities in Seshat and Chronos.

The experiment design is 80% training and 20% test split with R^2 as evaluation metrics. Basically, the higher the score, the more variance can be explained. The regression is performed with linear modeling (linear regression) and non-linear modeling (random forest regression).

Table 1
Regression of target variables with all features in Chronos-trv and Seshat-pca. The best results are marked in bold.

dataset	algorithm	target	R^2
chronos-trv	linear regression	GOV	0.747
chronos-trv	linear regression	SCALE	**0.744**
chronos-trv	linear regression	HIER	**0.829**
chronos-trv	random forest	GOV	0.721
chronos-trv	random forest	SCALE	**0.804**
chronos-trv	random forest	HIER	**0.860**
seshat-pca	linear regression	GOV	**0.789**
seshat-pca	linear regression	SCALE	0.719
seshat-pca	linear regression	HIER	0.796
seshat-pca	random forest	GOV	**0.853**
seshat-pca	random forest	SCALE	0.697
seshat-pca	random forest	HIER	0.798

The results, reported in Table 1, show that TRVs explain more variability than OHE-PCA only in the case of SCALE and HIER dependent variables, both with linear and non-linear modeling. In particular, the difference is greater with non-linear modeling, It is interesting to note that OHE-PCA yields better models of the GOV dependent variable. This suggests that TRVs are better predictors of data expressed as numerical scales, while OHE-PCA are better in the case of prediction of data that is originally one-hot encoded.

These findings rise the question of how are the independent variables distributed with respect to the dependent variables. Figure 2 reports the distribution of all variables as histograms, and reveals that the GOV dependent variable has a multimodal distribution, while SCALE and HIER have log-normal distribution. This confirms that TRVs are best suited for predicting homogeneous and continuous variables.

These results also raise the question of how much each independent variable contributes to the prediction. Ablation studies were done to answer this question. Given that the distribution of the variables is not normal, Spearman correlations were used for the analysis.

Results, reported in Table 2, show at least four interesting phenomena:

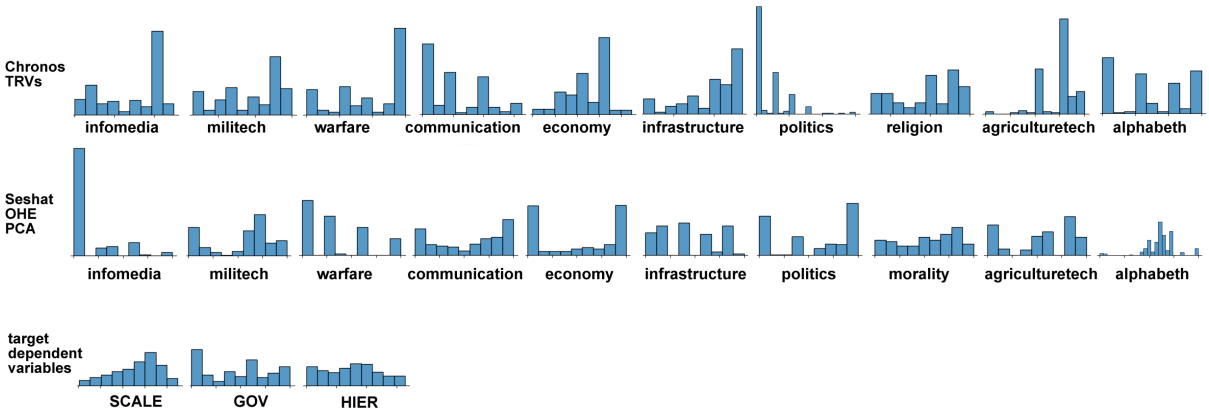

Figure 2: Histograms representing the distribution of all variables.

Table 2
Spearman correlations between the features in the two datasets and the target variables. *=p-value <0.005; **=p-value <0.001. The best correlations are marked in bold.

feature	dimensionality	SCALE	GOV	HIER
chronos-trv-warfare	5	0.857**	0.765**	0.811**
chronos-trv-infomedia	5	0.763**	**0.793****	0.815**
chronos-trv-infrastructure	6	**0.806****	0.766**	**0.822****
chronos-trv-communication	4	**0.774****	**0.732****	**0.831****
chronos-trv-economy	5	**0.826****	**0.746****	**0.801****
chronos-trv-militech	6	**0.748****	**0.679****	**0.739****
chronos-trv-religion	6	**0.764****	**0.766****	**0.760****
chronos-trv-alphabeth	3	0.596**	0.499**	0.653**
chronos-trv-agritech	4	**0.312****	**0.298****	**0.318****
chronos-trv-politics	4	-0.013**	-0.000**	-0.020**
seshat-pca-warfare	38	**0.868****	**0.778****	**0.831****
seshat-pca-infomedia	9	0.763**	0.790**	**0.826****
seshat-pca-infrastructure	9	0.735**	**0.796****	0.809**
seshat-pca-communication	3	0.755**	0.721**	0.814**
seshat-pca-economy	6	0.741**	0.653**	0.719**
seshat-pca-militech	5	0.745**	0.654**	0.714**
seshat-pca-morality	7	0.672**	0.695**	0.731**
seshat-pca-alphabeth	5	**0.705****	**0.753****	**0.754****
seshat-pca-agritech	3	0.119*	0.200*	0.175
seshat-pca-politics	3	**0.204***	**0.199**	**0.274****

- dimensionality has an impact on the predictive power of aggregated variables, but the improvement is not linear, as evidenced by the difference between in the warfare variable, where *seshat-warfare* has just a slightly higher performance with a much larger dimensionality;
- the p-values of the TRVs in the Chronos dataset are always below 0.001, unlike in Seshat-pca;
- the TRVs of the Chronos dataset yield most of the best correlations despite the dimensionality of the time-resolved variables is, by average, lower than in the OHE-PCA variables in Seshat;
- religion type is more informative than morality.

It is interesting to note that, given roughly the same dimensionality, TRVs tend to have higher correlation strength to the target variables, like in the case of *chronos-economy*, *chronos-militech*, *chronos-communication*, *chronos-religion* and *chronos-agritech*. This means that TRVs are able to summarize the information better than OHE-PCA on historical data.

3.4. Correlation Analysis and Interpretation

A correlation analysis is performed on the Chronos-TRV and Seshat-PCA datasets to answer RQ2 and test whether TRVs are useful for the interpretation of historical data. This is a qualitative evaluation from data visualization.

seshat-pca	infomedia	techmilitary	warfare	communication	economy	infrastructure	politics	morality	agritech	alphabeth
infomedia										
techmilitary			0.63	0.77	0.85	0.71	0.75	0.80	0.66	
warfare		0.63		0.71	0.70	0.66	0.59	0.70	0.58	
communication		0.77	0.71		0.84	0.77	0.73	0.79	0.64	
economy		0.85	0.70	0.84		0.77	0.80	0.85	0.69	
infrastructure		0.71	0.66	0.77	0.77		0.68	0.80	0.70	
politics		0.75	0.59	0.73	0.80	0.68		0.80	0.68	
morality		0.80	0.70	0.79	0.85	0.80	0.80		0.87	
agritech		0.66	0.58	0.64	0.69	0.70	0.68	0.87		
alphabeth										

chronos-trv	infomedia	techmilitary	warfare	communication	economy	infrastructure	politics	religion	agritech	alphabeth
infomedia		0.75	0.86	0.71	0.80	0.84		0.82		0.62
techmilitary	0.75		0.86	0.62	0.87	0.65		0.72		0.64
warfare	0.86	0.86		0.66	0.86	0.80		0.81		0.69
communication	0.71	0.62	0.66		0.67	0.61		0.64		
economy	0.80	0.87	0.86	0.67		0.78		0.76		0.62
infrastructure	0.84	0.65	0.80	0.61	0.78			0.79		0.64
politics										
religion	0.82	0.72	0.81	0.64	0.76	0.79				0.52
agritech										
alphabeth	0.62	0.64	0.69		0.62	0.64		0.52		

Figure 3: Filtered spearman correlation matrices of the independent variables in the two datasets.

First of all, correlation matrices are extracted from the two datasets. Only strong and significative correlations between the independent variables are kept, filtering out self-correlations; the dependent variables (GOV, SCALE, HIER); the dimensions with ρ between -0.5 and 0.5 and with $p\text{-}value > 0.001$. These matrices, reported in Figure 3, reveal that

- there are no inverse correlations;
- correlation strength is similar with the two information compression techniques;
- different variables were filtered out in the two datasets.

In particular, in Seshat-PCA the infomedia and alphabeth dimensions are missing, while in Chronos-TRV there are no politics and agritech variables. It is clear that the two datasets, compressed with different techniques, are showing different knowledge representations and narratives.

The interpretation of Seshat-PCA is not straightforward. PCA transforms correlated OHE variables into a set of linearly uncorrelated principal components, hence strong correlations between components means they contribute similarly to the direction of maximum variance, likely having similar loadings on the first principal component. In other words, high correlation between OHE-PCA transformed variables means they are pulling together in a similar direction and, in the context of polities, this can be interpreted as different variables contributing to a specific social goal.

On the contrary, the interpretation of TRVs in Chronos is more transparent since stages of increasing complexity are represented in each dimension. Under this perspective, high correlation between two or more TRV-compressed variables means that they tend to increase complexity together.

In order to have a better overview of the relations between dimensions, the matrices were visualized as correlation graphs, reported in Figure 4.

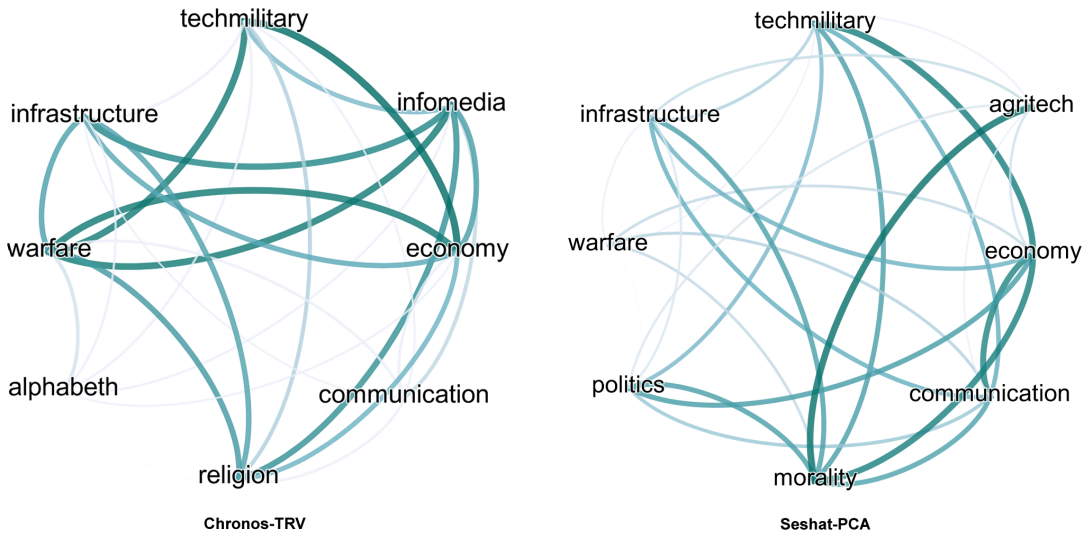

Figure 4: Correlation graphs of Chronos and Seshat-PCA. Only variables with $\rho > 0.5$ and p-value <0.001 are displayed. The thickness and darkness of the edges is proportional to the correlation strength.

The Seshat-PCA graph shows a general model of dimensions that cooperate towards a goal. For example this graph presents a strong pattern of interplay between military technology, economy, morality and agriculture. An explanation for this comes from the Structural Demographic Theory (SDT) [25], which states that war, distributed wealth, agricultural productivity and strong morality in religion can contribute to gain and keep societal stability.

Chronos-TRV shows a different pattern, only partially overlapping with the one in Seshat-PCA. Chronos-TRV shows strong correlations between the stages of increasing complexity in military technology, economy, wafare tactics, information management and – with less strength – infrastructure control and religion. This rich pattern can be interpreted in the light of the "Ratchet Effect" theory of cultural transmission [26]. This theory posits that societal advancements, once achieved, are rarely lost. As societies develop, they build upon existing knowledge and technology, leading to a cumulative process that increase complexity, and the Chronos-TRV graph plots the strength of dimensions involved in the process. New military technologies often lead to economic growth, with war or deterrence; a strong economy can boost military research as well as fund warfare [27]. At the same time effective warfare tactics often rely on superior information management and intelligence. Moreover, religion can provide a unifying ideology, legitimizing state authority to invest effort in warfare or infrastructure development, and this process is supported by technologies for information management, like media [28].

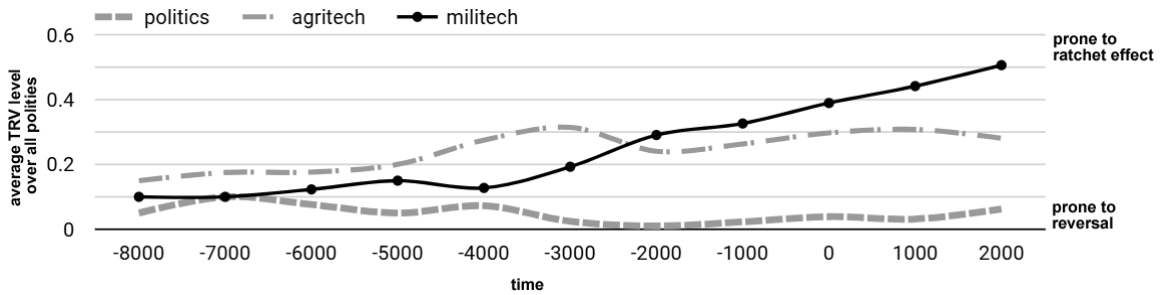

Figure 5: Timeline of Chronos-TRV variables comparing the average advancements in politics, agritech and militech. Values are averaged over all polities.

However, not all the variables considered here present a ratchet effect pattern. The politics and agritech variables are missing in the Chronos-TRV graph, because they have a correlation coefficient below 0.5. This means that advancements in these dimensions are more subject to reversal than others.

A timeline (Figure 5) comparing the average advancements in politics, agritech and militech from the Chronos-TRV dataset confirms this.

In practice, the more a line tends upward, the more the variable is likely to follow a ratchet effect. Societies that evolve in military technology, economy, warfare, information management or control on the infrastructures often gain a competitive advantage and tend to increase complexity also in the other dimensions in order to keep it.

4. Conclusion and Future Work

This paper addressed the issue of Knowledge Discovery and Representation for historical data in the field of Digital Humanities, dealing in particular with the problem of data integrity. The experiments presented in this paper suggest findings in the following areas:

- **Data Compression**. From a technical point of view, TRV is an interpretable information compression technique, based on sequences and scholarly literature. It is suitable to explain variance of numerical variables with non-linear modeling, and to visualize complex phenomena through time. In order to have advantages over existing automated compression techniques, it is possible to use specific prompts to produce TRV-annotated data with LLMs. The suggested experimental design is to manually annotate a small dataset as a ground truth, generate TRVs-annotated data with LLMs, then testing the error in the generated data comparing it to the ground truth.
- **Hypothesis testing in Digital Humanities**. Results reported in Table 2 show that religion type is more informative than morality in the prediction of social complexity. This is an hypothesis testing not just for different data mining techniques, but also for different theories. Consider that the variable *seshat-pca-morality* derives its categories from a systematic assessment of literature about Axial age theory [29], while the *chronos-trv-religion* variable lines up religion types by historical appearance, and can be subsumed under the umbrella of evolutionary theories of religion [30]. This means that, in principle, theories can be encoded and tested with compression techniques, and TRVs are suitable to encode theories that can provide scales or meaningful orders between categories. Moreover, order is not necessarily related to time. For example it is possible to use a cold-to-hot scale to encode categorical variables such as climate or geography into interpretable TRVs.
- **Knowledge Discovery**. TRV encoding showed that politics and agritech variables do not follow a ratchet effect pattern. This suggests that it is necessary to distinguish between *evolutive variables*, that are linear over time (and more likely to show a ratchet effect), and *adaptive variables*, that show non-linear behavior over time (and more subject to reversal). Evolutive variables, like advances in military technology or economy, determine a competitive advantage that a society can no longer be given up. No known human society that adopted iron weapons ever abandoned them for bronze weapons. On the contrary, political systems evolve and involve easily. For example the Roman republic and the egalitarian societies of the neolithic introduced collective assemblies and population representatives, but in the following times these were replaced by many polities governed by sole rulers.
- **Model Explanation**. Results of the correlation analysis suggest that TRVs can have a positive impact on interpretability, because we know in advance the semantics of each level of the scale and its relation to the following one. It is important to stress the fact that correlation is not a cause-effect relation and sequentiality is not causality (*post hoc ergo propter hoc fallacy*) but rather an interplay of Granger causality. Hence it is possible to associate different semantic relations to the steps of the scale. For example a semantic relation between the copper and the bronze level of the militech variable is *enablement*. In other words the ability to work copper enables the ability to work bronze, that is an alloy of copper and tin. In the same way the ability to manipulate symbols like seal stamps paves the way to the use of cuneiform writings for administrative information management. Under this perspective, the interpolations provided by TRVs can be useful to

advance hypotheses on unknown things. For example we do not know what religion there was in Çatalhöyük, but it was something in-between a cult of the ancestors and a cult of fertility. The same for Göbekli Tepe, where TRVs predict a transition between a shamanic cult of the dead and a cult of ancestors. Moreover, in a network structure it is possible to generalize sematic relations to the edges between different dimensions: for example infrastructure control can *facilitate the circulation of* economic capital, and media can *improve the dissemination of* religious practices. Hence, it is possible to build a knowledge graph from historical data compressed with TRVs.

The present work paves the way towards a new model of compressed knowledge representation for historical data that is at the same time interpretable and machine-readable. Future work in this direction includes:

- testing prompts to generate annotation of TRVs with LLMs;
- testing LLM-grounding to link TRV annotation to existing scholarly literature, that can be useful for digital libraries;
- testing the inter-annotator agreement in the use of TRVs, both between humans and between LLMs;
- experimenting with TRV step values encoding the sequence of events, like in the present paper, or both the sequence and the time between events;
- comparing TRVs against other compression techniques;
- applying TRVs in domains different than history.

This work presented a manual annotation of TRVs, whose purpose was to test their efficacy against OHE-PCA in a controlled setting. The dataset produced is available online in a Google sheet[1] among all the other datasets produced within the Chronos project for collaborative open science under Creative Common attribution, non-commercial share alike license.

Acknowledgments

This research was supported by the European Commission grant 101120657: European Lighthouse to Manifest Trustworthy and Green AI - ENFIELD.

This research employed data from the Seshat Databank (seshatdatabank.info) under Creative Commons Attribution Non-Commercial (CC By-NC SA) licensing.

References

[1] W. J. Frawley, G. Piatetsky-Shapiro, C. J. Matheus, Knowledge discovery in databases: An overview, AI magazine 13 (1992) 57–57.

[2] R. Roller, Theory-driven statistics for the digital humanities: Presenting pitfalls and a practical guide by the example of the reformation, Journal of Cultural Analytics 7 (2023).

[3] N. Horsley, What can a knowledge complexity approach reveal about big data and archival practice?, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 2246–2250.

[4] G. Demartini, K. Roitero, S. Mizzaro, Data bias management, Communications of the ACM 67 (2023) 28–32.

[5] D. L. Barbera, E. Maddalena, M. Soprano, K. Roitero, G. Demartini, D. Ceolin, D. Spina, S. Mizzaro, et al., Crowdsourced fact-checking: Does it actually work?, Information Processing & Management 61 (2024).

[6] E. Hyvönen, Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web 11 (2020) 187–193.

[1] https://docs.google.com/spreadsheets/d/1OW6CtmUudN3WTJ1VvWRZYZdTWVEjDJGns6Q8_I6EBwk/edit?usp=sharing

[7] K. Golub, Y.-H. Liu, Information and knowledge organisation in digital humanities: Global perspectives, Taylor & Francis, 2022.
[8] G. Di Nunzio, J. A. Dyble, F. Giachelle, S. Gialdroni, et al., Linking historical evidence to digital maps: The micoll map, in: CEUR WORKSHOP PROCEEDINGS, volume 3536, 2023, pp. 104–112.
[9] F. Giachelle, J. Dyble, G. M. Di Nunzio, S. Gialdroni, Exploring historical routes and waypoints with micoll digital map, in: International Conference on Theory and Practice of Digital Libraries, Springer, 2024, pp. 141–150.
[10] L. Davide, M. Rovera, F. Alfio, S. Tonelli, et al., Timeframe: Querying and visualizing event semantic frames in time, in: Proceedings of the First Workshop on Reference, Framing, and Perspective@ LREC-COLING 2024, ELRA and ICCL, 2024, pp. 13–17.
[11] A. Locaputo, B. Portelli, S. Magnani, E. Colombi, G. Serra, Ai for the restoration of ancient inscriptions: A computational linguistics perspective, in: Decoding Cultural Heritage: A Critical Dissection and Taxonomy of Human Creativity through Digital Tools, Springer, 2024, pp. 137–154.
[12] D. Firmani, F. Leotta, J. G. Mathew, J. Rossi, L. Balzotti, H. Song, D. Roman, R. Dautov, E. J. Husom, S. Sen, et al., Intend: Intent-based data operation in the computing continuum, in: CEUR WORKSHOP PROCEEDINGS, volume 3692, CEUR-WS, 2024, pp. 43–50.
[13] P. Turchin, H. Whitehouse, P. François, D. Hoyer, A. Alves, J. Baines, D. Baker, M. Bartokiak, J. Bates, J. Bennet, et al., An introduction to seshat: Global history databank, Journal of Cognitive Historiography 5 (2020) 115–123.
[14] A. W. Johnson, T. K. Earle, The evolution of human societies: from foraging group to agrarian state, Stanford University Press, 2000.
[15] H. M. Johnson, Religion in social change and social evolution, Sociological Inquiry 49 (1979) 313–339.
[16] P. Turchin, H. Whitehouse, S. Gavrilets, D. Hoyer, P. François, J. S. Bennett, K. C. Feeney, P. Peregrine, G. Feinman, A. Korotayev, et al., Disentangling the evolutionary drivers of social complexity: A comprehensive test of hypotheses, Science Advances 8 (2022) eabn3517.
[17] I. Ul Haq, I. Gondal, P. Vamplew, S. Brown, Categorical features transformation with compact one-hot encoder for fraud detection in distributed environment, in: Data Mining: 16th Australasian Conference, AusDM 2018, Bahrurst, NSW, Australia, November 28–30, 2018, Revised Selected Papers 16, Springer, 2019, pp. 69–80.
[18] M. Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, E. Tuzhilina, Principal component analysis, Nature Reviews Methods Primers 2 (2022) 100.
[19] G. Chiarot, C. Silvestri, Time series compression survey, ACM Computing Surveys 55 (2023) 1–32.
[20] F. Celli, D. Mingazov, Knowledge extraction from llms for scalable historical data annotation, Electronics 13 (2024) 4990.
[21] N. Rane, S. Choudhary, J. Rane, Gemini versus chatgpt: applications, performance, architecture, capabilities, and implementation, Performance, Architecture, Capabilities, and Implementation (February 13, 2024) (2024).
[22] F. Celli, The wiki music dataset: A tool for computational analysis of popular music, arXiv preprint arXiv:1908.10275 (2019).
[23] P. Turchin, R. Brennan, T. Currie, K. Feeney, P. Francois, D. Hoyer, J. Manning, A. Marciniak, D. Mullins, A. Palmisano, et al., Seshat: The global history databank, Cliodynamics 6 (2015).
[24] P. Savage, Additional robustness analyses confirm that complex societies precede moralizing gods throughout world history, Nature Ecology & Evolution Community (blog) 5 (2019).
[25] J. A. Goldstone, Demographic structural theory: 25 years on, Cliodynamics 8 (2017).
[26] J. C. Landon, World History and the Eonic Effect: Civilization, Darwinism, Xlibris Corporation, 2010.
[27] F. Clifford, F. Baum Christopher, The effect of war on economic growth, Cato Journal 40 (2020).
[28] J. Müller, T. N. Friemel, Dynamics of digital media use in religious communities—a theoretical model, Religions 15 (2024) 762.
[29] D. A. Mullins, D. Hoyer, C. Collins, T. Currie, K. Feeney, P. François, P. E. Savage, H. Whitehouse, P. Turchin, A systematic assessment of "axial age" proposals using global comparative historical

evidence, American Sociological Review 83 (2018) 596–626.
[30] P. Boyer, B. Bergstrom, Evolutionary perspectives on religion, Annual review of anthropology 37 (2008) 111–130.