Crime and Society: A Multivariate Approach

Luca Marchesi* Luca Orlando† Tommaso Zipoli‡

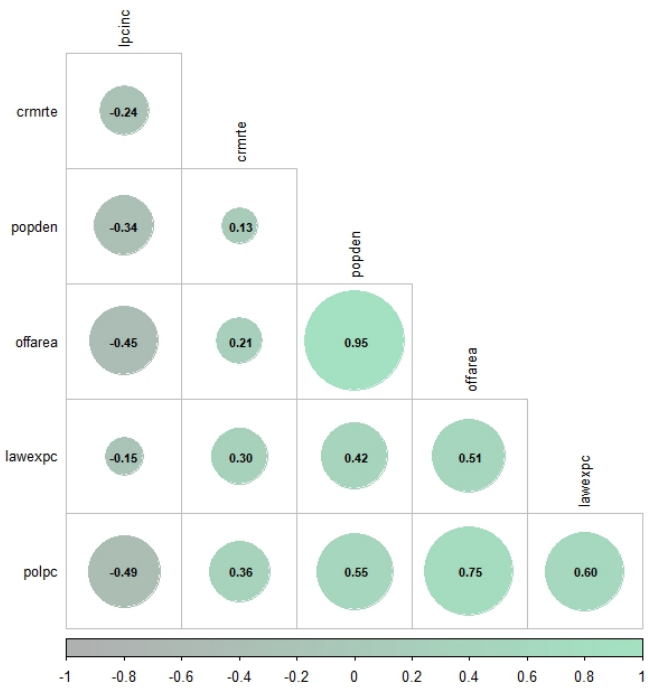

Figure 1: Correlation matrix

Introduction and Data Processing

Variable	Explanation
`crmrte`	Crimes per 1,000 population
`popden`	People per square mile
`lawexpc`	Law enforcement expenditure per capita
`offarea`	Police officers per square mile
`polpc`	Police officers per capita
`lpcinc`	Log of per capita income

Table 1: Selected Variables and Their Explanations

The objective of this analysis is to examine the relationship between urban crime rates and socio-economic characteristics at the city level in the United States. Specifically, we focus on how crime correlates with factors such as population density, income, and law enforcement attributes. Understanding these associations can provide insights into the structural and institutional conditions that may influence crime patterns in urban areas.

The dataset used for this study includes 46 observations, each corresponding to a U.S. city, and contains 19 variables related to reported crimes in 1982. For the purposes of our analysis, we selected a subset of variables guided by two principles. First, we prioritized *relative* measures over absolute ones—for example, crime rate instead of the total number of crimes, and population density rather than raw population figures—on the grounds that relative metrics are more informative for comparative analysis across cities of different sizes. Second, we favored *level* variables over their logarithmic transformations to minimize unnecessary manipulation—with the exception of income, for which only the log-transformed measure was available. All variables were subsequently standardized to ensure comparability and to prevent any single feature from disproportionately influencing the results due to differences in scale.

Figure 1 illustrates the correlation matrix of the selected variables, computed on standardized data. At a preliminary stage, we observe a strong correlation among law enforcement indicators, as well as between these and population density. There is also a moderate correlation with income per capita. In contrast, crime rates do not exhibit notably strong correlations with any of the other variables, suggesting a more complex or indirect relationship with the socio-economic and institutional features considered.

Principal Component Analysis (PCA)

Methodological Approach

In this section, we present the methodological framework employed in our analysis, focusing on the dimensionality reduction techniques that facilitate interpretation without compromising statistical rigor. Given the high-dimensional nature of the original dataset, we adopt principal component analysis (PCA) as a core technique to extract the most informative structures within the data.

PCA allows us to transform a set of d potentially correlated variables $\mathbf{X} = [X_1, X_2, \ldots, X_d]$ into a new set of uncorrelated components $\mathbf{W} = [W_1, W_2, \ldots, W_d]$, known as principal components, which successively maximize the variance explained. These components are orthogonal by construction and ordered according to the amount of total variance they capture. To formalize this, the principal components must satisfy the following properties:

*LMEC student, no. 1178522
†LMEC student, no. 1171758
‡LMEC student, no. 1176258

(a) $\mathrm{Cov}(W_k, W_{k'}) = 0 \quad$ for all $k \neq k'$,

(b) $\mathrm{Var}(W_1) > \mathrm{Var}(W_2) > \ldots > \mathrm{Var}(W_d)$,

(c) $\sum_{k=1}^{d} \mathrm{Var}(W_k) = \sum_{i=1}^{d} \mathrm{Var}(X_i)$.

To derive the principal components, we employ the spectral decomposition of the covariance matrix Σ, where $\Sigma = \Gamma\Lambda\Gamma^\top$. Here, Λ is a diagonal matrix with strictly positive and descending eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_d > 0$, and Γ is an orthogonal matrix whose columns η_k represent the eigenvectors (loadings) associated with each λ_k.

The k-th principal component is then defined as:

$$W_k = \sum_{i=1}^{d} \eta_{ki} X_i = \eta_k^\top \mathbf{X},$$

with η_{ki} indicating the loading of variable X_i on component W_k.

To ensure identifiability and mutual orthogonality of components, the eigenvectors must satisfy the following conditions:

$$\sum_{k=1}^{d} \eta_{ki}^2 = 1 \quad \rightarrow \text{ to ensure uniqueness,}$$

$$\sum_{k=1}^{d} \eta_{ki}\eta_{kj} = 0 \quad \text{for } j \neq i \quad \rightarrow \text{ to ensure orthogonality.}$$

This approach enables us to retain the components that explain the largest share of variance, reducing noise and redundancy while preserving the core informational content of the original variable space.

Selection of the Optimal Number of Components

A crucial step in PCA is determining the optimal number of principal components to retain. The goal is to preserve as much of the total variance as possible while reducing dimensionality for interpretability and efficiency. Several complementary criteria are employed in this selection:

- **Explained Variance Criterion:** The proportion of total variance explained by each component is computed, and a cumulative threshold (typically 70% or 80%) is used to decide how many components to retain.
- **Scree Plot:** A graphical tool plotting the eigenvalues λ_k in descending order. The "elbow point", where the slope of the curve levels off, is used as a heuristic cutoff.
- **Kaiser Criterion:** Retain only components with eigenvalues greater than 1, assuming standardized variables. This criterion filters out components that explain less variance than a single original variable.

In our analysis, we combine the explained variance threshold, the Kaiser criterion and the scree plot method to strike a balance between dimensionality reduction and informational completeness. This ensures that retained components are both statistically meaningful and interpretable in practical terms.

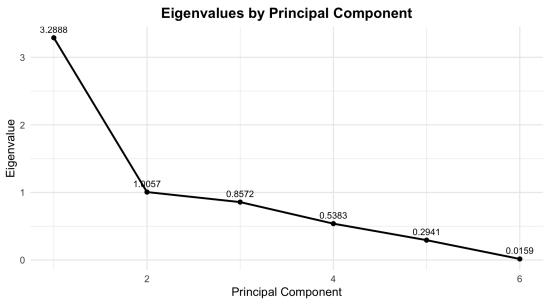

Figure 2: Relative variance explained by each principal component

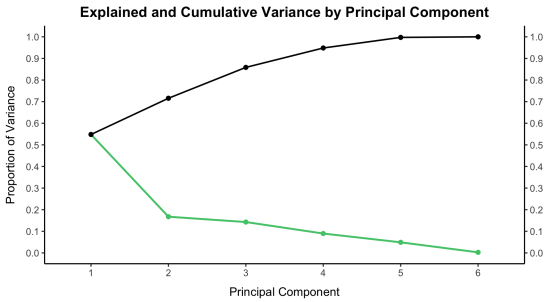

Figure 3: Cumulative variance explained by each principal component

Analysis

As already widely described in the methodology section, the primary aim of the PCA, as a multidimensional reduction technique, is to summarize the data without losing too much information with a linear combination of components. The incoming table (Table 2) presents the relative variance explained by each principal component and the cumulative percentage of the total variance.

Component	Variance	% Explained	Cumulative %
PC1	3.29	54.81	54.81
PC2	1.01	16.76	71.57
PC3	0.86	14.29	85.86
PC4	0.54	8.97	94.83
PC5	0.29	4.90	99.73
PC6	0.02	0.27	100.00

Table 2: PCA: Relative Variance and Cumulative Variance Explained

Furthermore, we leverage graphical visualizations of both the relative (Figure 2) and cumulative (Figure 3) variance explained by each principal component to determine the optimal number of components to retain. Additional meth-

ods considered include commonly used thresholds based on eigenvalues (e.g., 0.7 or 1) or cumulative variance (typically 70–80%).

Although the third principal component accounts for a non-negligible share of variance and would typically be retained under standard dimensionality reduction criteria, its inclusion significantly reduces the interpretability of the component structure. The sharp decline in explained variance between the first and second components, identifiable as a kink in the scree plot, already indicates a meaningful drop in marginal contribution. Adding the third component would complicate the interpretation without offering clear conceptual gains. Therefore, we opt to retain only the first two components, privileging clarity of interpretation over marginal increases in explained variance.

Variable	η_1^*	η_2^*
crmrte	0.418	0.825
popden	0.830	-0.422
lawexpc	0.679	0.188
polpc	0.871	0.112
offarea	0.933	-0.298
lpcinc	-0.580	-0.101

Table 3: Loadings of Variables on the First Two Principal Components

Variable	$\tilde{\eta}_1$	$\tilde{\eta}_2$
crmrte	0.127	0.820
popden	0.253	-0.420
lawexpc	0.206	0.186
polpc	0.265	0.112
offarea	0.284	-0.296
lpcinc	-0.176	-0.101

Table 4: Standardized Loadings

The presented tables (Table 3, Table 4) display respectively the component loadings η_{jk}^*, which are a rescaled measure of how much a variable j contribute to the k-th principal component, and the standardized loadings $\tilde{\eta}_{ik}$, which are used to compute the individual score coefficient on a particular component. So as to be more specific the component loadings are an extremely valuable tool to inspect the meaning of the results obtained using this technique as rescaling the relative contribution of the variables on the components allows us to inflate the weights for the variables that account for the most of the variance and deflate the ones for the less explicative components. Such rescaling is obtained from the η_{jk}, which are obtained form the resolution of the constraint matrix algebra problem presented above and weighted as shown:

$$\eta_{jk}^* = \eta_{jk}\sqrt{\lambda_k} \quad \text{for } k = 1, \ldots, d;\ j = 1, \ldots, d$$

After this rescaling, we are able to interpret the component loadings η_{jk}^* as correlations between variables and components.

To complement the interpretation of the component loadings, and before proceeding with the comments attached to the results, we now turn to the computation of the component scores, which represent the projection of each individual observation onto the principal components. When working with standardized data, the score of an individual on the k-th principal component is computed as a linear combination of the standardized variables $x_1, x_2, \ldots, x_d$, using the eigenvector elements η_{ik} for $i = 1, \ldots, d$ as weights:

$$w_k = \eta_{1k}x_1 + \eta_{2k}x_2 + \cdots + \eta_{dk}x_d$$

The variance of this component score equals the corresponding eigenvalue: $\text{Var}(w_k) = \lambda_k$.

However, it is often convenient to standardize the component scores to have unit variance. This is achieved by dividing each score by the square root of the associated eigenvalue:

$$\tilde{w}_k = \frac{w_k}{\sqrt{\lambda_k}} \quad \Rightarrow \quad \text{Var}(\tilde{w}_k) = 1$$

Accordingly, the standardized component score can be expressed as:

$$\tilde{w}_k = \tilde{\eta}_{1k}x_1 + \tilde{\eta}_{2k}x_2 + \cdots + \tilde{\eta}_{dk}x_d$$

where the coefficients $\tilde{\eta}_{ik}$, known as *component score coefficients*, are given by:

$$\tilde{\eta}_{ik} = \frac{\eta_{ik}}{\sqrt{\lambda_k}}$$

These standardized loadings η_{ik} for $k = 1, 2$ are shown in Table 4 and are used to compute the observations scores for the first two principal components (Figure 5) discussed in the following paragraph.

To economically interpret the retained principal components, we examine the loadings of each variable on the first two components to understand their structural meaning within the dataset (Table 3). The first principal component (PC1) is characterized by strong and positive loadings on variables such as *offarea*, *polpc*, *popden*, and *lawexpc*, and a significant negative loading on *lpcinc*. This configuration suggests that PC1 captures a structural dimension related to urban security infrastructure and population concentration. The high values of this component correspond to municipalities that are densely populated and exhibit high levels of policing and law enforcement expenditure. This security dimension, associated with densely populated areas, also correlates negatively with per capita income, suggesting that the component identifies relatively poorer urban districts where institutional presence is strong despite lower individual wealth. The second principal component (PC2) loads heavily on *crmrte*, indicating that it reflects the relative intensity of crime per capita, rather than absolute crime figures. The strong positive loading on *crmrte* and negative loading on *popden* imply that this component differentiates between high-crime, low-density areas and more

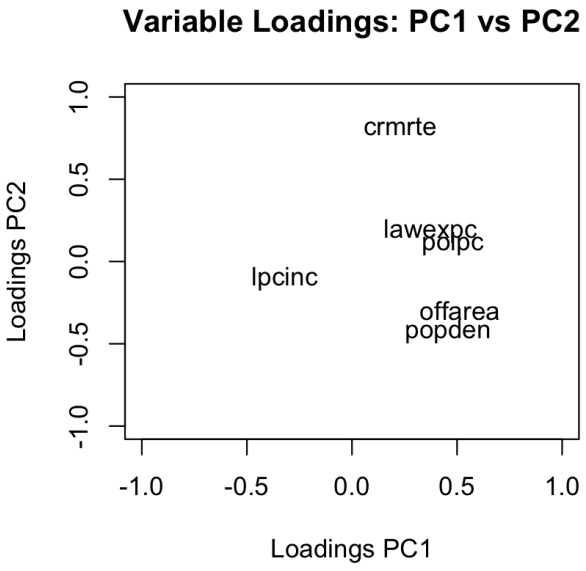

Figure 4: Variable loadings on the first two principal components

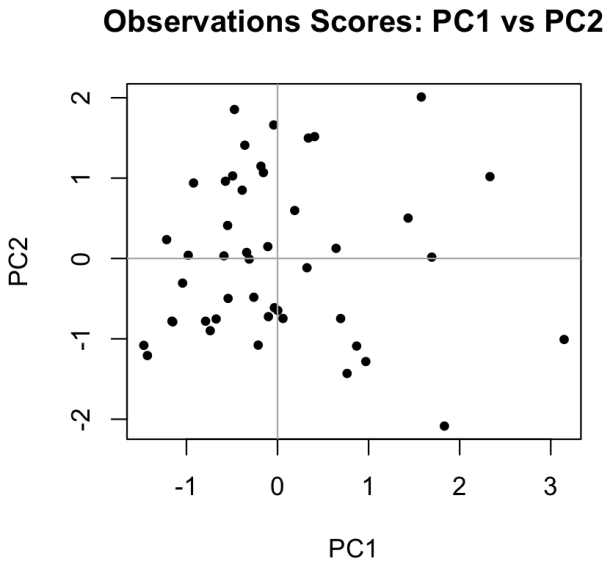

Figure 5: Observations scores on the first two principal components

compact urban settings with lower recorded crime rates. Other variables, such as *offarea* and *lpcinc*, contribute less strongly to this axis and do not reverse its interpretation.

Figure 4 shows how the principal components relate to the original variables, visually confirming our earlier interpretation. The first component clearly separates highly urbanized and institutionally secured districts, while the second captures variation in crime rates in opposition to population density. Figure 5, instead, shows the distribution of the observations in the new component space, obtained by computing the city score from the standardized loadings shown in Table 4. The component scores computed for each observation highlight a clear opposition between the two structural dimensions identified through PCA. Observations with high scores on PC1 and low scores on PC2 typically correspond to densely populated, low-income municipalities with high levels of institutional presence (e.g., policing and law enforcement expenditures), but with a relatively lower per capita crime rate as an endogenous result of such presence; conversely, observations with high scores on PC2 and low to moderate (in a few cases) on PC1 tend to represent territories with lower population density and limited urban infrastructure, yet show a greater exposure to crime per inhabitant. These municipalities may be spatially larger or more peripheral, with weaker institutional coverage.

Factor Analysis

Methodological Approach

In this section, we present the methodological framework underlying the factor analysis (FA) performed in our study. Unlike Principal Component Analysis, which is purely descriptive, factor analysis is a model-based technique used to uncover latent structures that account for the covariances among observed variables.

The Factor Model

We assume that the observed variables $x_1, x_2, \ldots, x_d$ can be modeled as linear combinations of q unobserved latent factors $f_1, f_2, \ldots, f_q$, plus a specific noise component:

$$x_i = \mu_i + \lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{iq} f_q + u_i, \quad i = 1, \ldots, d$$

In matrix notation, the model can be written as:

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda \mathbf{f} + \mathbf{u}$$

where:

- $\mathbf{x}$ is the $d \times 1$ vector of observed variables,
- $\boldsymbol{\mu}$ is the corresponding mean vector,
- Λ is the $d \times q$ matrix of factor loadings,
- $\mathbf{f}$ is the $q \times 1$ vector of latent factors,
- $\mathbf{u}$ is the $d \times 1$ vector of specific (unique) components.

The underlying assumptions include:

- $E(\mathbf{f}) = 0$, $E(\mathbf{f}\mathbf{f}') = I$ (uncorrelated, standardized factors),
- $E(\mathbf{u}) = 0$, $E(\mathbf{u}\mathbf{u}') = \Psi$ (diagonal matrix),
- $\mathrm{Cov}(\mathbf{f}, \mathbf{u}) = 0$ (factors and unique terms are uncorrelated).

The total variance-covariance structure is thus expressed as:

$$\Sigma = \Lambda\Lambda' + \Psi$$

Model Estimation and Factor Rotation

The parameters Λ and Ψ are estimated by maximizing the likelihood function under the assumption of multivariate normality. We adopt the Maximum Likelihood Estimation (MLE) method, which provides consistent and efficient estimates and allows us to formally test the goodness of fit of the factor model.

Once the initial factor solution is obtained, we apply both orthogonal (Varimax) and oblique (Oblimin) rotations to facilitate interpretation. Orthogonal rotation maintains factor independence, whereas oblique rotation allows correlation among latent factors—arguably more realistic in socio-economic contexts. Examining the factor correlation matrix, reported below, will offer stronger empirical grounds for supporting the oblique rotation.

Selection of the Optimal Number of Factors

Determining the appropriate number of factors q is critical to model identification and interpretability. We employ a combination of statistical and substantive criteria:

- **Likelihood-Ratio Test:** The chi-square test statistic compares the fit of models with differing numbers of factors. A significant reduction in the test statistic supports the inclusion of an additional factor.
- **Inspection of the residuals:** Small residuals, obtained by subtracting the replicated correlation matrix from the observed one, indicate that the factor model reproduces the observed correlations accurately.
- **Communalities:** The sum of the communalities represents the variance explained by the model; high communalities suggest that the factor model successfully accounts for most of the variance in the observed variables, supporting the model's adequacy.

Analysis

The primary aim of the factor model is to find one or more latent variables which could provide a satisfactory explanation of the original variables. We first estimated a 1-factor model (see Methodology), but the associated test statistic was not satisfactory based on the likelihood-ratio test ($\chi^2 = 53.168$, $df = 9$, $p < 0.001$), indicating that an additional factor was needed to improve the model's adherence to its underlying assumptions. The 2-factor model, by contrast, yielded a statistically significant result and an adequate fit ($\chi^2 = 4.735$, $df = 4$, $p = 0.3155$), satisfying both statistical and substantive criteria.

Factors	Chi-square	Degrees of freedom	p-value
1	53.168	9	0.0000
2	4.735	4	0.3155

Table 5: Wald test statistics for different factor models

The factor loadings displayed in the following table (Table 6) represent the correlation of each variable with every factor (similarly to the loadings presented in the PCA):

Variable	$\hat{\lambda}_{1i}$	$\hat{\lambda}_{2i}$
crmrte	0.228	0.302
popden	0.936	-0.302
lawexpc	0.530	0.293
polpc	0.784	0.600
offarea	0.996	
lpcinc	-0.459	-0.216

Table 6: Factor loadings for each variable in the 2-factor model

The first factor ($\hat{\lambda}_{1i}$) displays strong positive loadings on *offarea* (0.996), *popden* (0.936), *polpc* (0.784), and *lawexpc* (0.530), and a moderate negative loading on *lpcinc* (-0.459). This configuration suggests that Factor 1 captures a structural dimension of urban institutional intensity, combining spatial compactness, high population density, and strong investment in policing and law enforcement. The negative sign on *lpcinc* implies that this institutional intensity is particularly prominent in lower-income urban districts.

The second factor ($\hat{\lambda}_{2i}$) shows its strongest loading on *polpc* (0.600), followed by *crmrte* (0.302), *lawexpc* (0.293), and a negative loading on *popden* (-0.302). This indicates that Factor 2 reflects crime-related expenditure and exposure, with a partial inverse relationship to population density. In other words, this factor may be capturing districts that are less dense but relatively more exposed to crime, where policing efforts remain substantial.

An important diagnostic result emerges from the analysis of communalities (Table 8): although more than 60% of the variance is explained by the 2-factor model, both *crmrte* and *lpcinc* display consistently low communalities across the two- and three-factor solutions. This indicates that these variables are only marginally explained by the latent factors extracted and do not align structurally with the dimensions captured by the rest of the indicators. In substantive terms, this suggests that *crmrte* and *lpcinc*

Variable	Varimax F1	Varimax F2	Quartimax F1	Quartimax F2	Oblimin F1	Oblimin F2
crmrte	0.063	0.374	0.156	0.345	-0.125	0.444
popden	0.969	0.165	0.979	-0.087	1.037	-0.091
lawexpc	0.335	0.505	0.452	0.403	0.121	0.523
polpc	0.418	0.895	0.631	0.759	0.011	0.981
offarea	0.908	0.413	0.983	0.169	0.835	0.235
lpcinc	-0.308	-0.404	-0.400	-0.312	-0.143	-0.406

Table 7: Rotated factor loadings (2-factor solution)

are shaped by dynamics that differ from those driving the remaining variables, such as urban density, administrative presence, or public expenditure.

Variable	Communality
crmrte	0.144
popden	0.966
lawexpc	0.367
polpc	0.975
offarea	0.995
lpcinc	0.258

Table 8: Estimated communalities for each variable in the 2-factor model

In particular, the case of *crmrte* is illustrative. When attempting to discriminate across observations based on crime rate scores, no clear clustering or structural separation emerges, unlike what is observed with institutional or demographic variables. This empirical flatness reinforces the notion that crime rates are not tightly linked to the dominant latent dimensions in the dataset. Instead, they likely reflect idiosyncratic or context-specific factors—such as local law enforcement practices, social cohesion, or informal governance dynamics—that are not captured by the current model specification.

Likewise, per capita income (*lpcinc*) likely embodies broader economic trajectories that are only weakly correlated with the institutional or spatial features represented by the extracted factors. The limited gain in explanatory power after the inclusion of a third factor further confirms that these variables remain conceptually orthogonal to the core structure of the factor model. Rather than being statistical anomalies, they appear to represent substantively distinct aspects of the data-generating process, whose modeling may require additional variables or a framework capable of capturing both shared and unique sources of variance.

In order to improve interpretability, we performed both orthogonal and oblique rotations on the extracted two-factor solution (Table 7). Specifically, we applied the Varimax and Oblimin criteria and compared the resulting loadings. While both rotations yield broadly similar structural patterns, the underlying assumptions and interpretive implications differ. From a technical standpoint, the orthogonal rotation assumes that the extracted factors are uncorrelated by design; conversely, Oblimin relaxes the independence constraint and allows for correlation between latent factors. Empirically, both rotations identify a first factor strongly positively loaded by *popden*, *offarea* and a second factor by *polpc* and *lawexpc* alongside a more moderate weight on *crmrte*. In the Varimax solution, this structure is clearly separable: the first axis captures urban compactness and institutional density, while the second reflects crime-related institutional intensity. The moderate to strong negative loading of lpcinc on both factors reinforces the idea that lower income levels tend to characterize areas with both high density and policing intensity. In contrast, the Oblimin rotation highlights a stronger loading of *polpc* on the second factor (0.981), a still dominant role of *popden* on the first (1.037) and more cross-loadings on variables such as *lawexpc* and *crmrte*. This suggests a partially overlapping structure in which institutional presence and urban density are not fully independent but rather intertwined dimensions of the same socio-territorial process. We question the robustness of such claim by inspecting the correlation matrix of the latent factors, which reveals a value of 0.62. This strongly supports the decision to apply an oblique rotation. Thus, The Oblimin rotation may better reflect the conceptual complexity of urban environments, where administrative presence, crime exposure, and demographic structure are rarely orthogonal. However, the orthogonal rotation allows us to gain clarity and interpretability in the underlying latent factors. Summarizing, the main takeaway from this dual analysis is that while both PCA and FA identify broadly similar latent dimensions, one associated with urban density and institutional infrastructure, the other with crime-related patterns, the factor model introduces a conceptual distinction between shared and unique variance. This distinction allows us to separate variance common to multiple indicators from variance specific to each variable, thus offering a more precise structural interpretation. By explicitly modeling unique components, the factor model reinforces the evidence of multidimensionality already suggested by the PCA, while also confirming the substantive plausibility of the two-factor solution.

Cluster Analysis

Methodological Approach

Cluster analysis is an *unsupervised* learning technique aimed at partitioning data into homogeneous groups—or clusters—without prior knowledge of either the membership of each observation or the number of clusters. The broad criterion is that observations within each cluster should be more *similar* to each other than to those in different clusters. In the context of our analysis, similarity is measured by Euclidean distance.

Clustering algorithms can be broadly divided into hierarchical and non-hierarchical approaches. **Hierarchical (agglomerative)** algorithms start by treating each observation as its own singleton cluster and progressively merge clusters according to a specific criterion, producing a k-cluster solution for any value of k. The clustering pattern can be visualized using a tree-like object called a *dendrogram*. In this analysis, we primarily rely on *Ward's method*, which forms clusters by minimizing the within-cluster variance at each step:

$$W = \sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - \bar{x}_j\|_2^2$$

where $j = 1, \ldots, k$ indexes the clusters and $\bar{x}_j$ denotes the d-dimensional centroid of cluster j. Other hierarchical methods, such as single linkage, complete linkage, and the centroid method, will be referenced for comparison purposes.

Non-hierarchical algorithms (often referred to as k-means) begin with a fixed, predetermined number of clusters and then relocate observations to minimize an objective function, generally linked to the within-cluster variance. We will employ the *Lloyd–Forgy algorithm*, which consists of the following steps:

1. Randomly select k observations as initial cluster centroids.
2. Assign each of the remaining observations to the cluster with the nearest centroid. Once all observations have been assigned, recalculate the centroids and compute the within-cluster variance (WCV).
3. Using the updated centroids, repeat step 2. At each iteration, the WCV is expected to decrease. The algorithm stops when it converges, i.e., when the cluster assignments no longer change or, equivalently, when the WCV settles down.

Selection of the Optimal Number of Clusters

A key issue in cluster analysis is selecting the optimal number of clusters. This can be addressed in multiple ways:

- *Heuristic methods*. These include visual inspection of the dendrogram and cluster profiling.
- *Variance-based methods*. These involve selecting the number of clusters corresponding to an 'elbow' in the scree plot (a plot of within-cluster variance against the number of clusters k). At the elbow point, further increasing k yields diminishing returns in terms of variance reduction, indicating limited additional informational gain.
- *Silhouette analysis*. This method evaluates how similar an observation is to its own cluster compared to those in other clusters. The silhouette coefficient ranges between -1 and 1, where 1 indicates a clear cluster assignment and -1 suggests possible misclassification.[1]

We adopt elbow analysis for the Ward method and silhouette analysis for k-means.

Cluster	crmrte	popden	lawexpc
1	0.3005	1.2420	0.7012
2	-0.1060	-0.4384	-0.2475
Cluster	**polpc**	**offarea**	**lpcinc**
1	1.2380	1.3929	-0.7797
2	-0.4369	-0.4916	0.2752

Table 9: Cluster centroids (Ward) for $k = 2$

Analysis

Cluster profiling

The tree-like structure of groupings (or *dendrogram*) obtained via the Ward method (Figure 6) exhibits a clear bifurcation at $k = 2$. These two clusters are primarily separated by a factor we refer to as **"exposure to crime,"** which accounts for approximately 40% of the total variance. Cities in the first cluster exhibit higher population

[1] The silhouette coefficient is defined as:

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

where $s(i)$, the silhouette value for observation i, is given by:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

with:

- $a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, j \neq i} d(i,j)$, the average dissimilarity between point i and all other points j in the same cluster C_i,
- $b(i) = \min_{C \neq C_i} \left(\frac{1}{|C|} \sum_{j \in C} d(i,j) \right)$, the minimum average dissimilarity between point i and all points in other clusters $C \neq C_i$,
- $d(i,j)$ is the distance between points i and j, typically measured using Euclidean distance,
- $|C_i|$ is the number of points in the same cluster as i, and $|C|$ is the number of points in any other cluster.

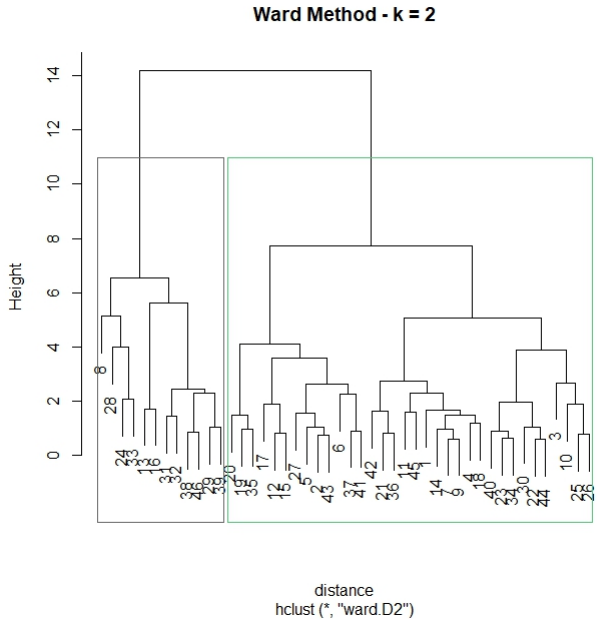

Figure 6: Dendrogram for Ward-based clustering

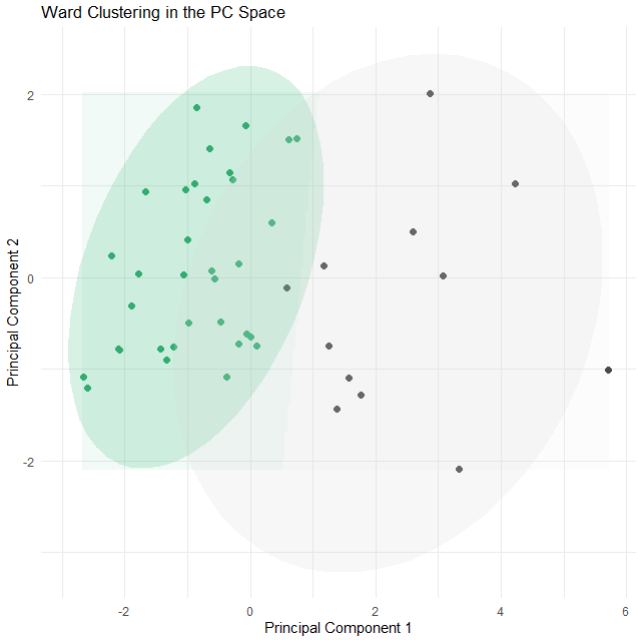

Figure 7: Ward clustering in the PC space

density and lower income per capita—factors that appear to contribute to higher crime rates (see Table 9). Correspondingly, these cities are characterized by stronger law enforcement, measured by both the relative number of officers and security expenditure (although the latter shows a smaller gap between the two clusters). This increased law enforcement may exert a feedback effect on the crime rate—an issue of endogeneity that we explore in the final section.

In contrast, cities in the second cluster are less exposed to crime, likely due to their lower population density and higher per capita income. These cities indeed exhibit lower crime rates, even though law enforcement presence is significantly reduced. The results are further supported by the Lloyd–Forgy algorithm: with $k = 2$, the cluster assignments are identical to those of the Ward method, except for a single observation, resulting in only negligible differences in the computed centroids.

To further investigate the nature of the hidden factor—previously referred to as exposure to crime—two figures provide additional insight. Figure 8 projects the Ward-based clustering into the space spanned by the first two principal components. The fictitious separation line between the two clusters is vertical, indicating that the division is strongly aligned with the first principal component. This component is characterized by strong positive loadings for population density and police presence (both per capita and per area), and a significant negative loading for income. Remarkably, crime rate itself contributes only marginally, suggesting that our concept of "exposure to crime" is not strictly aligned with observed crime rates.

Figure 7 displays the same clusters in various bivariate spaces to highlight the contribution of individual variables.

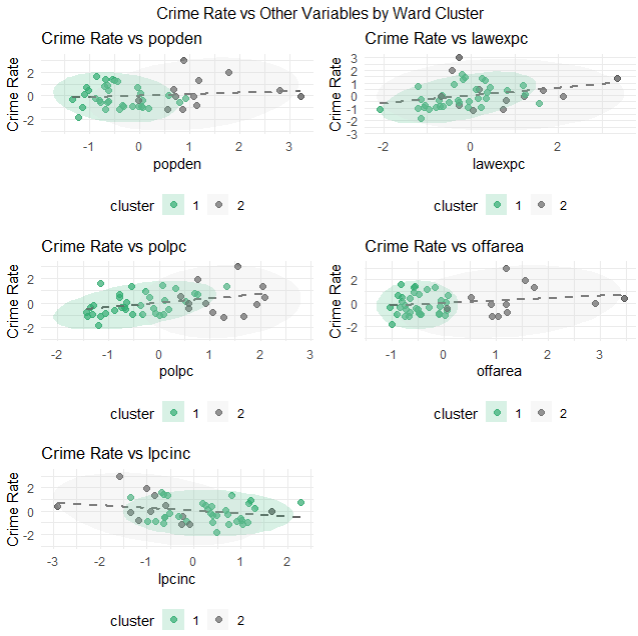

Figure 8: Crime rate against other variables by cluster

Again, population density and number of officers emerge as the most influential factors. Notably, there is no clear boundary between high- and low-crime cities. If crime rate were a principal discriminating factor, we would expect to observe a stark vertical separation in these plots—with one cluster consistently appearing above the other.

Diagnostic and Robustness

In this section, we first motivate our choice of the number of clusters, then assess the stability and validity of our findings by comparing them with the results of alternative clustering methods. As previously mentioned, visual inspection of the dendrogram naturally suggests $k = 2$. To corroborate this choice, the scree plot shows that further partitioning does not lead to significant increases in explained variance (i.e., the total variance minus the within-group variance; see the "elbow" at $k = 2$ in Figure 9). For non-hierarchical clustering, the silhouette plot (Figure 10) consistently points to the same optimal number of clusters.

The results obtained via complete linkage mirror those from the Ward method one-to-one, reinforcing the robustness and interpretability of the underlying structure. In contrast, the single linkage and centroid methods are unsuitable for this analysis. Both are highly sensitive to outliers—each being affected by a distinct extreme observation—which results in groupings that lack substantive interpretability.

As for k-means clustering, the results are robust to different initialization values (as confirmed through multiple runs). Moreover, the *Hartigan–Wong* algorithm converges to the same grouping.[2]

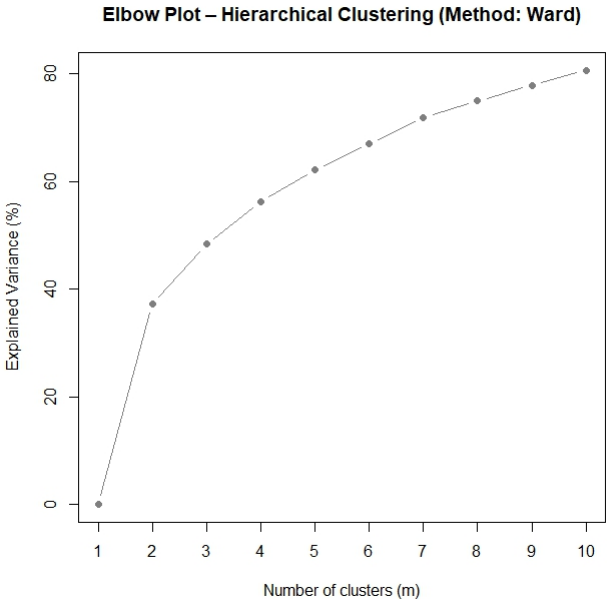

Figure 9: Analysis of the explained variance

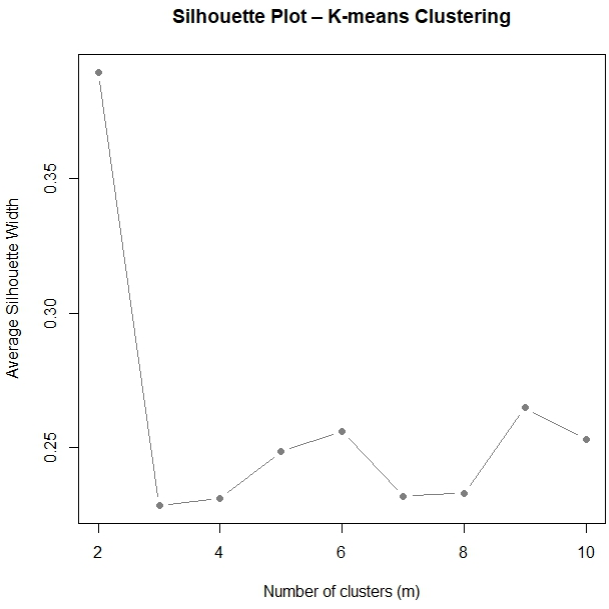

Figure 10: Silhouette analysis

Conclusion

The combined application of PCA and factor analysis allowed for a reduction in dimensionality and the identification of latent structures underpinning the observed correlations. In particular, both methods isolate a first dimension broadly associated with urban density and institutional infrastructure, and a second axis more closely tied to criminal activity and enforcement patterns.

Despite general alignment between the two techniques, relevant differences emerge. Most notably, factor analysis reveals a significant portion of unexplained variance for variables such as *crmrte* and *lpcinc*. This result suggests that neither crime rates nor per capita income follow the same latent trajectories as other institutional or demographic variables. In other words, these outcomes are not reducible to common structural or socioeconomic dimensions captured by the factor model.

This insight is analytically relevant. It indicates that crime and individual wealth may evolve independently of

[2]The Hartigan–Wong algorithm is similar to the Lloyd–Forgy method but updates group assignments point-wise rather than in batches.

institutional configurations or demographic density, thus requiring dedicated modeling approaches beyond the scope of shared latent structures. While our analysis remains limited in scope and exploratory in nature, this finding underscores the importance of not over-relying on dimensional reduction when the underlying phenomena may be driven by distinct and context-specific mechanisms.

The results of the cluster analysis reinforce these conclusions. The grouping of observations identifies two structurally distinct clusters based on crime exposure—a hidden discriminant feature heavily loaded on population density, law enforcement, and average income. However, this separation does not extend clearly to observed crime rates, which appear to be distributed more heterogeneously across the clusters.

One possible explanation lies in the idea that crime dynamics are not primarily determined by the structural dimensions uncovered through exploratory methods. Instead, they may reflect localized effects that require more targeted, context-aware approaches to be fully understood—approaches that go beyond the structure of the analyzed dataset.

A competing (or potentially complementary) explanation for the observed weak alignment between crime rate and socioeconomic variables is the potential endogeneity of key predictors. Cities with high exposure to crime may increase law enforcement intensity, which in turn may lead to a subsequent reduction in crime. The strong alignment between population density, income per capita, and police-force indicators supports this thesis. However, to fully assess its validity, the dataset should be expanded to include additional socioeconomic features, such as immigration rates or unemployment levels.

Finally, it would be valuable to explore whether regional variations in institutional and demographic factors further influence the relationship between crime rates and socioeconomic conditions. Differences in regional policies, policing strategies, and local economic conditions could yield diverse outcomes, highlighting the importance of considering geographical context in future investigations of crime dynamics and socioeconomic interactions. A region-specific analysis could reveal patterns that are obscured in broader national-level data, providing more targeted insights for policy-making. It is crucial to note that our analysis relies on a single cross-sectional dataset from 1982, and this snapshot of the variables severely limits the ability to capture dynamic processes and causal relationships. Moreover, as briefly mentioned above, the absence of certain socioeconomic variables may restrict the scope of the latent structures uncovered. Extending the study to longitudinal data and incorporating a richer set of indicators would allow for a more comprehensive assessment of the stability, evolution, and policy relevance of the identified patterns.