UDC 519.2
DOI https://doi.org/10.24144/2616-7700.2025.47(2).199-206

Y. M. Okuniev

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv School of Economics,
Student, TA of Probability Theory and Statistics,
Master's Student
egorky96@gmail.com
ORCID: https://orcid.org/0009-0009-3695-490X

CLUSTERING OF REGIONS IN UKRAINE BASED ON STATE OF HIV/AIDS EPIDEMIC

It is a well-known fact that Ukraine is among the countries most affected by the HIV/AIDS epidemic. In this work, I fit a mixed linear model to cluster the regions of Ukraine based on the number of HIV/AIDS diagnoses and AIDS-related deaths. Using this model, I represent each region as a combination of stochastic vectors, each describing the probabilities of a region belonging to a particular class based on the normalized numbers of HIV/AIDS diagnoses and AIDS-related deaths in that region. I then construct a graph where each region is represented as a vertex, and each edge is weighted by the correlation between the corresponding class-belonging vectors. Finally, I perform Walktrap clustering to group regions according to the similarity of HIV/AIDS epidemic trends and interpret the resulting clusters. The aforementioned work is conducted as part of an actuarial study of the HIV/AIDS epidemic in Ukraine.

Keywords: Mixed linear models, Clustering analysis, Walktrap community detection, Weighted correlation networks, Network analysis of epidemics

1. Literature review. The HIV/AIDS epidemic remains one of the most pressing challenges for Ukraine's healthcare system [1,2]. According to international reports, Ukraine ranks among the countries with the highest rates of HIV/AIDS transmission in Europe and globally [3, 4]. Addressing this public health crisis requiers a formalized, data-driven framework for modeling the epidemic at the regional level.

Previous studies have extensively studied the HIV epidemic using molecular biology [5], phylodynamic approaches [6], and other data-driven methods. However, these works typically focus on regions with a high concentration of internally displaced persons or areas already known to have a significant epidemic burden, leaving other regions and the effectiveness of their preventive and anti-HIV interventions comparatively underexplored.

2. Introduction. This study focuses on analyzing the regions of Ukraine to identify patterns in the dynamics of the HIV/AIDS epidemic. The goal is to classify regions into groups according to their epidemiological risk levels and to interpret the resulting clusters in terms of shared epidemic trends. This research contributes to the ongoing effort to understand the spatial and temporal structure of the HIV/AIDS epidemic in Ukraine.

To achieve these objectives, the following steps were performed:

- Collection and preparation of regional data on HIV/AIDS incidence and AIDS-related mortality in Ukraine;
- standardization of the data;
- fitting multiple mixed linear models with varying numbers of latent states based on standardized indicators of HIV/AIDS diagnoses and AIDS-related deaths [7];

- selecting the models that best capture the key epidemiological characteristics of the epidemic;
- deriving class-belonging probability vectors for each region based on the fitted models;
- constructing a weighted correlation network in which vertices represent regions and edges reflect the correlational similarity of their epidemiological profiles;
- applying the Walktrap clustering algorithm to identify groups of regions with similar epidemic dynamics [8];
- interpreting the obtained clusters to reveal common regional traits and factors contributing to shared epidemic patterns.

3. Data preparation. For the purposes of this study, region-level data were obtained from the State Enterprise "Center of Public Health" of the Ministry of Health of Ukraine [9], covering the period from January 2014 to the present. The dataset provides monthly observations for each region and includes the following data:

- number of newly reported cases of HIV,
- number of newly reported cases of AIDS,
- number of AIDS-related deaths,
- total number of individuals with HIV/AIDS under medical supervision at the end of each month,
- corresponding rates per 100000 population.

The original data were stored in a format unsuitable for direct import into Microsoft Excel. For further processing, Optical Character Recognition (OCR) technology was applied using ABBYY software, which specializes in recognizing spreadsheet-like document structures that closely resembled the format of the source files.

The automated OCR process introduced several technical challenges. Specifically, the software occasionally misclassified the digits 0 and 1 as the letters "o" and "i," introduced sporadic data corruption, and misinterpreted center-aligned symbols as entries containing multiple spaces. To correct these issues, manual data cleaning was performed, including the removal of redundant spaces and systematic replacements such as $o \to 0$ and $i \to 1$.

It is important to emphasize that the available data reflect only officially documented cases of HIV infection and AIDS progression. The true prevalence of HIV infection remains uncertain due to underdiagnosis and incomplete reporting.

Throughout this study, the following notational conventions are used:

- G — name of group G being studied,
- N — increment in the size of a given group,
- T — total number of individuals in a given group,
- D — number of AIDS-related deaths,
- .100k — indicator expressed per 100000 population.

Additionally, the left subscript "t" denotes the time index. For instance, the total number of individuals living with HIV at time point 4 is represented as ${}_4HIV.T$.

For analytical consistency, several regions were excluded from the dataset: Ukraine (national aggregate), Sevastopol, Luhansk region, and Donetsk region. The national-level data were omitted as redundant for a region-specific analysis, while the excluded

regions lack consistent observations across the study period due to temporary occupation and limited data avalability and credibility.

Finally, transitions between HIV and AIDS stages (progressions and regressions) were not explicitly modeled. As all individuals with HIV/AIDS under medical supervision currently receive antiretroviral therapy (ART), which may revert AIDS to HIV and induce viral remission, such transitions are unobservable given the available empirical data.

4. Fitting mixed linear model. In this section, I describe the process of fitting mixed linear models used to categorize each region according to its HIV/AIDS epidemic state.

Specifically, three mixed linear models were estimated: one for the standardized linear trends in the number of HIV diagnoses, one for the standardized linear trends in the number of AIDS diagnoses, and one for the number of deaths related to AIDS.

To standardize the regional data, a z-transformation was applied, ensuring that, at each time point, the data have a mean of zero and a standard deviation of one. Normalization was performed on values of HIV.N.100k, AIDS.N.100k and AIDS.D.100k in order to avoid biases caused by populational simillarity between regions.

Each of these models captures a distinct aspect of the epidemic dynamics within a region:

- the number of HIV diagnoses reflects both the susceptibility of a region's population to HIV infection and the effectiveness of preventive campaigns aimed at reducing new cases and promoting early diagnosis;
- the number of AIDS diagnoses primarily represents the prevalence of late, yet still manageable, HIV cases;
- finally, the number of deaths caused by AIDS captures the most severe epidemic outcomes—cases diagnosed too late for effective medical treatment.

For example, regions that exhibit a moderate number of HIV diagnoses, but a low number of AIDS diagnoses and AIDS-related deaths can be considered relatively 'low risk'. Conversely, regions with moderate or high levels of AIDS diagnoses and AIDS-related deaths may require improvements in public awareness, early testing, and quality of anti-HIV campaigns.

To estimate the latent class mixture models, the `lcmm` package for R was used. The modeling procedure involves solving a likelihood maximization problem to identify the best-fitting model for a specified number of latent classes. In this study, given the richness of the dataset, models with one, two, and three classes were fitted. The destandardized mean trajectories of these models are illustrated in the following figures.

From the plots, it is evident that, regardless of the number of classes or the characteristics of individual classes, a decreasing trend is present, indicating the effectiveness of the anti-HIV campaigns.

Furthermore, for the trajectories of AIDS diagnoses and AIDS-related deaths, the addition of a third class did not produce a substantially new class compared to the two-class model.

To finalize the selection of the number of latent classes for the studied models, the Bayesian Information Criterion (BIC) values were compared. The BIC values for studied models are presented in Table 1:

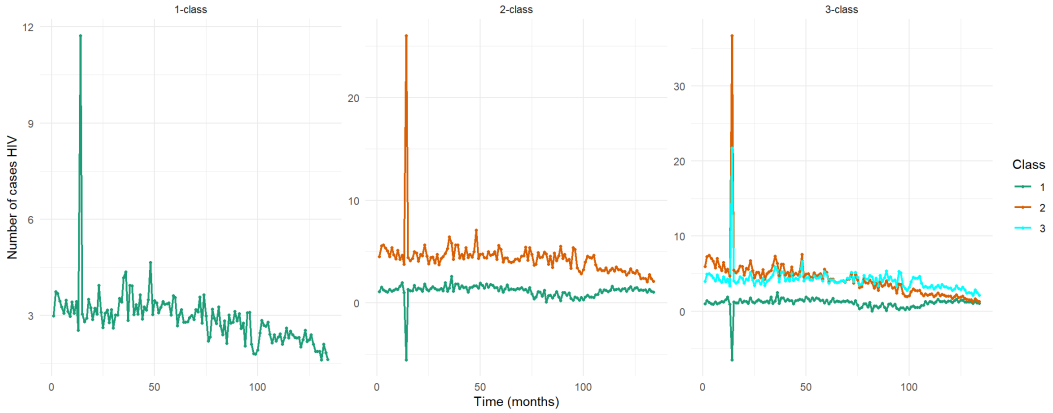

Figure 1. Expected trajectories of HIV.N.

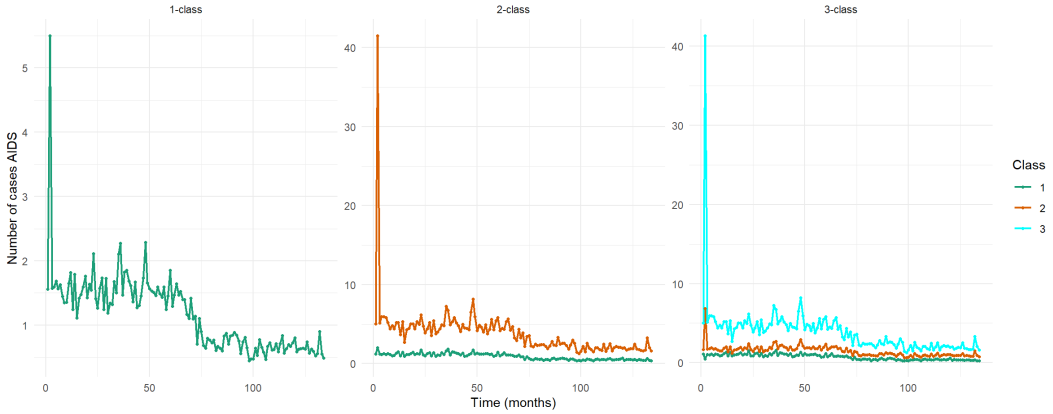

Figure 2. Expected trajectories of AIDS.N.

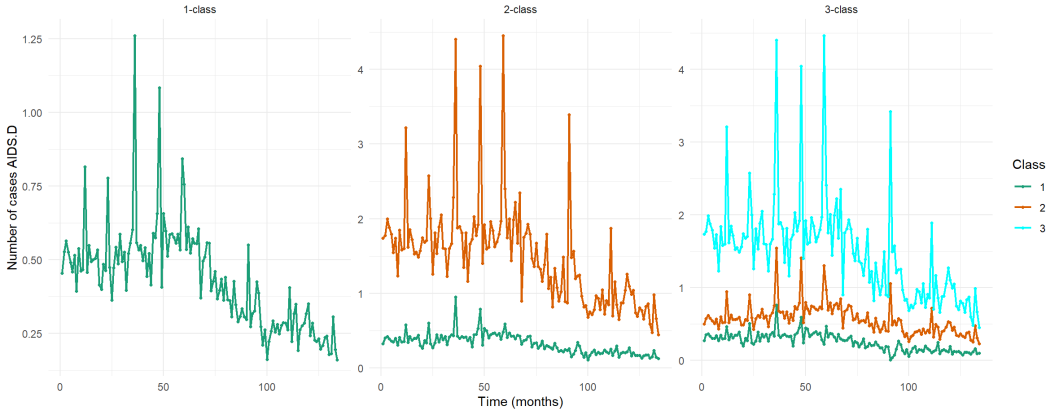

Figure 3. Expected trajectories of AIDS.D.

Therefore, 3-class models were chosen for HIV.N.100k and AIDS.D.100k, and 2-class model was chosen for AIDS.N.100k. The absence of improvement in interpretability for the three-class model of AIDS.N.100K is evident from the plot of expected trajectories. In this specification, two of the three classes substantially overlap, providing no additional meaningful differentiation and therefore offering

Table 1.

BIC values for latent class mixed models (HIV.N, AIDS.N, and AIDS.D).

Variable	**1-class**	**2-class**	**3-class**
HIV.N.100k	3435.553	3418.599	3412.978
AIDS.N.100k	4668.3	4660.506	4664.915
AIDS.D.100k	5497.002	5478.181	5477.191

limited explanatory value compared to the two-class model.

5. Building regional weighted correlation network.

Once each mixed linear model was fitted, three probability vectors were obtained for each region, corresponding to the likelihood of membership in each latent class based on HIV.N, AIDS.N, and AIDS.D, respectively.

To identify regions exhibiting similar epidemic patterns, a regional network was constructed in which each region represents a vertex, and each pair of vertices is connected by an edge weighted according to the correlation between the corresponding probability vectors.

To reduce informational clutter and eliminate redundant connections (e.g., weakly or negatively correlated regions), hard thresholding was applied. Specifically, a threshold τ was defined, and edges e with weight $\omega(e) < \tau$ were removed.

In order to choose appropriate τ, frequency histogram of values in correlation matrix was studied:

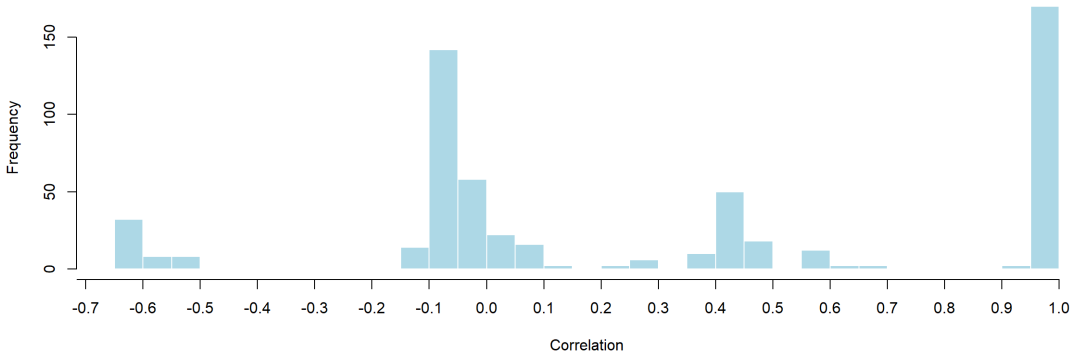

Figure 4. Histogram of values in correlation matrix

As evident from the histogram, most of the strong positive correlations are concentrated at values of 0.9 and above; therefore, a threshold of $\tau = 0.9$ was selected for subsequent analyses.

After hard thresholding, the Walktrap clustering algorithm was applied to group regions with similar epidemic trends. All network construction and clustering procedures were implemented using the `igraph` package in R.

The resulting network is illustrated in Figure 5.

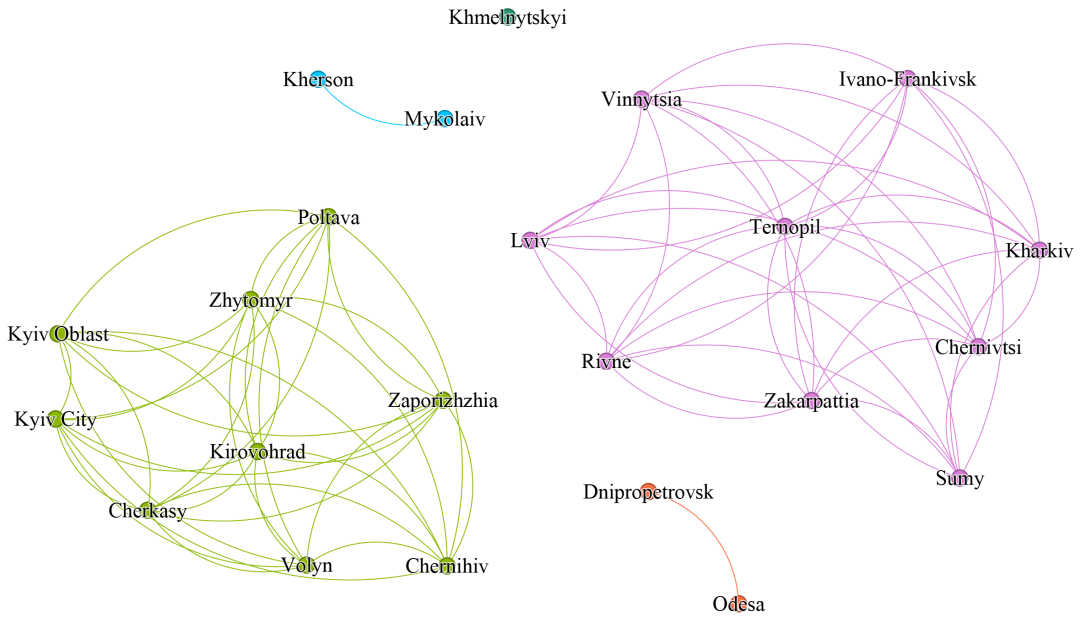

Figure 5. Epidemic regional clusters

With the network constructed, the resulting clusters were analyzed and interpreted in Table 2.

Table 2.

Cluster composition by model class assignments.

Cluster	Size	HIV.N Class	AIDS.N Class	AIDS.D Class
1	9	1	1	1
2	9	3	1	2
3	2	2	1	2
4	2	3	2	3
5	1	1	1	2

Cluster 1 includes regions with low numbers of HIV and AIDS diagnoses and low AIDS-related mortality. The apparent mildness of the epidemic may reflect under-testing or incomplete reporting rather than genuinely low prevalence.

Cluster 2 represents regions with moderate HIV incidence, low AIDS diagnoses, and moderate AIDS-related mortality (typically below one death per month). The relatively higher number of HIV diagnoses suggests broader testing coverage and a more accurate depiction of the epidemic situation.

Cluster 3 comprises regions with stable HIV incidence, low AIDS diagnoses, and moderate mortality. While epidemic trends appear under control, enhanced awareness campaigns could further reduce HIV transmission, so that those region follow common declining trend.

Cluster 4 covers regions historically exhibiting the highest HIV/AIDS burden — particularly Odesa and Dnipro. Despite a declining trend in HIV diagnoses, persistently high AIDS cases and deaths indicate that inadequate preventive and informational efforts remain key drivers of severity, consistent with previous studies [10].

Cluster 5 includes regions likely affected by underdiagnosis, though the epidemic is less severe than in Cluster 4.

6. Conclusions and Prospects for Further Research.

As a result of the conducted study, multiple mixed linear models were fitted to classify regions according to different fundamental aspects of the HIV/AIDS epidemic. Among the fitted models, the most appropriate were selected for subsequent analysis based on both interpretability and Bayesian Information Criterion (BIC) values. Using these models, a regional epidemiological network was constructed to identify regions that share similar epidemic patterns, and the resulting clusters of regions were thoroughly interpreted.

This work represents an important step toward the actuarial study of the epidemic at the regional level. Available epidemiological data are aggregated at the country level, which makes it challenging to infer detailed regional epidemic dynamics. Therefore, the clustering of regions and the estimation of correction coefficients are essential to capture local patterns and inform targeted interventions. This methodology provides a foundation for future research aimed at improving regional epidemic modeling and optimizing public health responses.

References

1. Kruglov, Y., Kobyshcha, Y., Salyuk, T., Varetska, O., Shakarishvili, A., & Saldanha, V. (2008). The most severe HIV epidemic in Europe: Ukraine's national HIV prevalence estimates for 2007. *Sexually Transmitted Infections,* 84(1), 37–41. https://doi.org/10.1136/sti.2008.031195
2. Barnett, T., Whiteside, A., Khodakevich, L., Kruglov, Y., & Steshenko, V. (2000). The HIV/AIDS epidemic in Ukraine: Its potential social and economic impact. *Social Science & Medicine,* 51(9), 1387–1403. https://doi.org/10.1016/s0277-9536(00)00104-0
3. Miranda, M. N. S., Pingarilho, M., Pimentel, V., Martins, M. R. O., Vandamme, A.-M., Bobkova, M., Böhm, M., Seguin-Devaux, C., Paredes, R., Rubio, R., Zazzi, M., Incardona, F., & Abecasis, A. (2021). Determinants of HIV-1 late presentation in patients followed in Europe. *Pathogens,* 10(7), 835. https://doi.org/10.3390/pathogens10070835
4. Parczewski, M., Gökengin, D., Sullivan, A., de Amo, J., Cairns, G., Bivol, S., Kuchukhidze, G., Vasylyev, M., & Rockstroh, J. K. (2025). Control of HIV across the WHO European region: Progress and remaining challenges. *The Lancet Regional Health–Europe,* 52, 1–14. https://doi.org/10.1016/j.lanepe.2025.101243
5. Vasylyeva, T. I., Liulchuk, M., Friedman, S. R., Sazonova, I., Faria, N. R., Katzourakis, A., ... & Magiorkinis, G. (2018). Molecular epidemiology reveals the role of war in the spread of HIV in Ukraine. *Proceedings of the National Academy of Sciences,* 115(5), 1051–1056. https://doi.org/10.1073/pnas.1701447115
6. Kovalenko, G., Yakovleva, A., Smyrnov, P., Redlinger, M., Tymets, O., Korobchuk, A., ... & Vasylyeva, T. I. (2023). Phylodynamics and migration data help describe HIV transmission dynamics in internally displaced people who inject drugs in Ukraine. *PNAS nexus,* 2(3), pgad008. https://doi.org/10.1093/pnasnexus/pgad008
7. Proust-Lima, C., Philipps, V., & Liquet, B. (2025). lcmm: Mixed models for longitudinal data and latent classes in R [R package]. Retrieved from https://cran.r-project.org/web/packages/lcmm/lcmm.pdf
8. Csardi, G., & Nepusz, T. (2025). igraph: Network analysis and visualization in R. Version 1.2.6 [R package]. Retrieved from https://igraph.org/r/html/1.2.6

9. State Enterprise "Center of Public Health" of the Ministry of Health of Ukraine. *Statistics on HIV/AIDS in Ukraine.* (2025). Retrieved from https://phc.org.ua/kontrol-zakhvoryuvan/vilsnid/statistika-z-vilsnidu
10. Neduzhko, O., Postnov, O., Perehinets, I., DeHovitz, J., Joseph, M., Odegaard, D., ... & Kiriazova, T. (2017). Factors associated with delayed enrollment in HIV medical care among HIV-positive individuals in Odessa Region, Ukraine. *Journal of the International Association of Providers of AIDS Care (JIAPAC),* 16(2), 168–173. https://doi.org/10.1177/2325957416686194

Окунєв Є. М. Кластеризація регіонів України на основі епідеміологічної ситуації ВІЛ/СНІД.

Відомо, що Україна належить до країн із найвищим рівнем поширення епідемії ВІЛ/СНІД. У цій роботі було побудовано лінійні змішані моделі, щоб кластеризувати регіони України за кількістю діагностованих випадків ВІЛ/СНІД та смертей, пов'язаних зі СНІДом. Кожен регіон представлено як комбінацію стохастичних векторів, що описують ймовірності належності регіону до певного класу на основі нормалізованих показників. Далі було побудовано граф, де регіони — це вершини, а ребра зважено за кореляцією відповідних векторів. Нарешті, за допомогою алгоритму Walktrap проведено кластеризацію регіонів за схожістю епідемічних тенденцій та інтерпретовано отримані кластери. Ця робота виконана в рамках актуарного дослідження епідемії ВІЛ/СНІД в Україні.

Ключові слова: Лінійні змішані моделі, кластерний аналіз, Walktrap кластеризація, зважені кореляційні мережі, мережевий епідеміологічний аналіз

Recived 11.09.2025