ЭКОНОМИЧЕСКАЯ И СОЦИАЛЬНАЯ ГЕОГРАФИЯ

ECONOMIC AND SOCIAL GEOGRAPHY

UDC 913 (574) https://doi.org/10.5281/zenodo.20786467

Celtikoglu S. [1,2,3], Cankurt M.[4], Reck A.[2], Oskenbayev Y.[5], Thomas T.[4]

[1]Independent Institute for Environmental Issues (UfU e.V.), Berlin, Germany
[2]Leibniz Institute of Agricultural Development in Transition Economies (IAMO), Halle (Saale), Germany
[3]Humboldt-Universität zu Berlin, Berlin, Germany
[4]North Carolina Agricultural and Technical State University, North Carolina, USA
[5]Narxoz University, Almaty, Kazakhstan

A DATA-DRIVEN SPATIAL TYPOLOGY OF AGRICULTURAL REGIONS IN KAZAKHSTAN: A PCA AND AFFINITY PROPAGATION APPROACH

Abstract. *This article develops a spatial typology of agricultural regions in Kazakhstan using principal component analysis and affinity propagation clustering. The study covers 16 territorial units and draws on regional data for 2015–2023, including demographic, climatic, geographic, and agricultural indicators. In the optimized four-variable PCA used for the main clustering stage, the first two principal components explained 99% of the variance. The analysis identified three interpretable clusters: lower-intensity dryland-oriented regions, high-capacity grain regions, and a distinct southern cluster with a differentiated production profile. The resulting typology provides a transparent regional classification that supports regional comparison and differentiated planning in Kazakhstan.*

Key words: *Kazakhstan, agricultural regionalization, spatial typology, principal component analysis, affinity propagation, regional planning, agricultural geography.*

Челтикоглу С. [1,2,3], Джанкурт М.[4], Рек А.[2], Оскенбаев Е.[5], Томас Т.[4]

[1]Независимый институт экологических проблем (UfU e.V.), Берлин, Германия
[2]Институт сельскохозяйственного развития в странах с переходной экономикой имени Лейбница (IAMO), Халле (Заале), Германия
[3]Гумбольдтский университет в Берлине, Берлин, Германия
[4]Государственный сельскохозяйственный и технический университет Северной Каролины, Северная Каролина, США
[5]Университет Нархоз, Алматы, Казахстан

ПРОСТРАНСТВЕННАЯ ТИПОЛОГИЯ СЕЛЬСКОХОЗЯЙСТВЕННЫХ РЕГИОНОВ КАЗАХСТАНА, ОСНОВАННАЯ НА ДАННЫХ: ПОДХОД С ИСПОЛЬЗОВАНИЕМ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ И РАСПРОСТРАНЕНИЯ СХОДСТВА

Аннотация. *В статье разработана пространственная типология сельскохозяйственных регионов Казахстана с использованием метода анализа главных компонентов и кластеризации методом распространения сходства. Исследование охватывает 16 территориальных единиц и основано на региональных данных за 2015–2023 годы, включая демографические, климатические, географические и сельскохозяйственные показатели. В оптимизированном четырехфакторном анализе главных компонентов, использованном на этапе основной кластеризации, первые два главных компонента объяснили 99% дисперсии. Анализ выявил три интерпретируемых кластера: регионы с низкой интенсивностью засушливых земель, регионы с высокой производительностью зерновых культур и отдельный южный кластер с дифференцированным профилем производства. Полученная типология обеспечивает*

прозрачную региональную классификацию, которая поддерживает региональное сравнение и дифференцированное планирование в Казахстане.

Ключевые слова: *Казахстан, сельскохозяйственное районирование, географическая типология, анализ главных компонентов, группировка по сходству, региональное планирование, география сельского хозяйства.*

Introduction and problem statement. Agricultural planning in large countries is fundamentally a spatial problem. When regional differences in production structure, environmental conditions, and settlement patterns are collapsed into national averages, planning instruments become less effective because they are applied to territories that do not share the same constraints or development potentials [1,2,8]. From a regional geographic perspective, the central challenge is therefore not only to describe territorial variation, but to identify spatially meaningful regional groupings that can support differentiated resource planning and allocation.

Study of the problem. Existing approaches to agricultural regionalization broadly fall into two categories. The first includes expert-based agro-ecological frameworks, such as AEZ and GAEZ, which provide valuable benchmarks for broad-scale land suitability and production potential [6]. The second includes data-driven regional typologies that use multivariate statistics and clustering to identify similarities among territorial units on the basis of observed characteristics [12,13,16]. The first approach is strong in biophysical standardization, whereas the second is better suited to detecting subnational heterogeneity in realized agricultural structure. For geographically extensive countries with strong internal variation, the two approaches should be understood as complementary rather than competing.

Kazakhstan provides a compelling case for such an analysis. As one of the world's largest countries, it contains major contrasts between northern grain-producing regions, arid and semi-arid dryland territories, and more diversified southern production systems. Although these broad contrasts are widely recognized, a transparent and reproducible region-level typology that integrates agricultural, climatic, geographic, and demographic information remains limited in the national literature. Existing studies are often sector-specific, environmentally narrow, or geographically restricted to selected parts of the country [3,15].

The aim and objectives of the work. The aim of this study is to develop a data-driven spatial typology of agricultural regions in Kazakhstan using principal component analysis and affinity propagation clustering. The specific objectives are: (1) to assemble a comparable regional dataset covering demographic, climatic, geographic, and agricultural indicators for 2015–2023; (2) to identify interpretable regional clusters based on the main dimensions of agricultural differentiation; and (3) to assess how the resulting typology can support regional comparison and differentiated planning discussions.

This article addresses that gap by developing a data-driven spatial typology of Kazakhstan's agricultural regions using principal component analysis (PCA) and affinity propagation (AP) clustering. The objective is not to propose a new agro-ecological theory, but to construct an empirically grounded regional classification that can support comparison across territories and inform regionally differentiated planning. In this sense, the article uses geographic methods to address practical problems of spatial differentiation, regional comparison, and territorial planning.

The article makes three specific contributions. First, it provides a transparent region-level classification of agricultural heterogeneity across Kazakhstan. Second, it demonstrates how a parsimonious PCA–AP workflow can be used to generate interpretable regional clusters in a data-constrained setting. Third, it shows how these clusters can serve as a spatial baseline for differentiated planning discussions in relation to grain-system concentration, production intensity, and broader territorial variation. More broadly, the study shows how geographically informed typologies can convert familiar regional contrasts into an operational framework for comparison and planning. The object of the study is the regional agricultural geography of

Kazakhstan as represented by 16 territorial units under the stable pre-2022 administrative configuration. The subject of the study is the spatial differentiation of agricultural production structures and their broader demographic, climatic, and geographic context as captured through the selected indicator set.

Materials and methods. This study develops a region-level spatial typology of agricultural production structures in Kazakhstan. The clustering sample consists of 16 territorial units drawn from the stable pre-2022 reporting geography: 13 regions plus the three republican cities of Astana, Almaty, and Shymkent. Mangystau is excluded from the clustering sample because agriculture occupies a comparatively marginal position in its regional economic structure and because its inclusion would contribute limited interpretive value to a production-oriented typology while increasing structural asymmetry within the analytical sample. The exclusion is therefore methodological rather than territorial: Mangystau is not omitted from the geography of Kazakhstan as such, but from the clustering model used to derive comparable agricultural production profiles. Newly established regions created in 2022 (Abai, Ulytau, and Zhetysu) are presented descriptively where relevant but are not included in the clustering model so that the analytical sample remains temporally consistent over 2015–2023.

The analytical objective is to derive an information-dense representation of regional agricultural heterogeneity from multiple correlated indicators and to identify territorially meaningful clusters without imposing conventional administrative groupings on the results. The workflow combines data harmonization, dimensionality reduction, clustering, and post hoc interpretation of the resulting regional patterns.

Data. The empirical basis of the study is an annual region-level dataset covering the period 2015–2023. For the purposes of clustering, these data are not used to model temporal trajectories directly; instead, they are harmonized and aggregated into comparable regional profiles that capture average structural characteristics over the observation period. The inferential logic of the study is therefore cross-sectional rather than longitudinal. Twelve indicators were assembled to capture key dimensions of regional variation, including agricultural scale, production structure, demographic context, and selected climatic and geographic conditions. Definitions, units, and sources are reported in Appendix A, Table A2.

The dataset combines official agricultural and demographic statistics from KazStat with climatic and geographic information obtained from Kazhydromet, NASA, FAO, and OECD-related sources.

Variables, units, and preprocessing. All indicators were harmonized in terms of definition, unit, temporal aggregation, spatial aggregation, and transformation. The observation window is 2015–2023, and the clustering inputs were constructed from region-level means or totals, as appropriate. Units were standardized across the dataset, including wheat production (10^3 t), precipitation (mm/year), mean daily sunshine duration (h/day), temperature (°C), agricultural land area (ha), and production value (million KZT). This standardization was carried through to all figures and tables to avoid ambiguity in interpretation.

The raw data underwent a structured preprocessing pipeline. First, missing values within each region's 2015–2023 series were imputed using forward-fill and linear interpolation, never across regions. Second, potential outliers were identified using the Tukey rule (1.5 × IQR), and variables were winsorized at p1–p99 on the region-year panel used to construct the clustering sample (16 regions × 9 years = 144 observations before missingness). Third, all clustering inputs were standardized as z-scores prior to PCA, while min–max normalization was used only for map visualization.

Treatment of post-2022 administrative changes. To preserve temporal comparability, the clustering analysis uses the stable territorial configuration that was consistently reported over most of the 2015–2023 period. The post-2022 administrative units Abai, Ulytau, and Zhetysu are therefore not included in the clustering model. Where these units appear in descriptive appendix materials, they are shown only for contextual reference and are not used in estimating the PCA or affinity propagation solution.

Analytical framework. The analytical design follows the ISAC framework—Identification, Segmentation, and Characterization—as a structured way of moving from regional data assembly to interpretable territorial groupings [18]. In the Identification stage, candidate variables were screened, harmonized, and reduced to a set of twelve indicators suitable for region-level comparison. In the Segmentation stage, dimensionality reduction was conducted through principal component analysis, followed by affinity propagation clustering. In the Characterization stage, the resulting clusters were interpreted in relation to their agricultural structure, environmental context, and planning relevance.

PCA was used to reduce multicollinearity, preserve the main variance structure of the dataset, and generate a lower-dimensional feature space suitable for clustering. Indicators were winsorized and standardized before PCA. The main specification uses the first two principal components (PC1–PC2) because they offer a parsimonious and interpretable representation of regional variation. As sensitivity checks, additional PCA spaces based on PC1–PC3 and PC1–PC4 were also examined.

Affinity propagation was then applied to the PCA scores. Similarity between regions was defined using negative squared Euclidean distance, $s(i,k) = -||x_i - x_k||^2$, where x_i and x_k denote the PCA coordinates of regions i and k. Although affinity propagation does not require the analyst to pre-specify the number of clusters, the number of resulting exemplars is sensitive to the preference parameter. We therefore treat cluster number as data-guided but analyst-conditioned rather than fully automatic.

In the main specification, the preference parameter was set to -10 because this value produced the most stable and substantively interpretable three-cluster solution across repeated runs and alternative PCA spaces. The damping parameter was set to 0.90, the maximum number of iterations to 1000, and the convergence criterion to 50 iterations with a fixed random seed. This parameterization was chosen to prioritize stability, interpretability, and geographical readability at the regional scale.

Unless otherwise noted, clustering is performed in the PC1–PC2 space derived from an optimized four-variable PCA including annual production value, annual wheat production, agricultural land area, and number of agricultural enterprises. This reduced specification is intentionally production-centered: it is designed to capture the core structure of regional agricultural output and enterprise organization in a parsimonious form. The broader twelve-variable PCA is retained for contextual interpretation and visualization, but it does not define the main cluster solution. Accordingly, the resulting clusters should be interpreted as production-oriented regional groupings that are subsequently discussed in relation to broader territorial conditions, rather than as direct agro-environmental classes.

Justification of methods. The choice of PCA and affinity propagation reflects the geometry of the data and the exploratory purpose of the study. High-dimensional agricultural datasets typically contain correlated indicators, and in such settings the discriminatory value of distance measures tends to decline—a problem widely described as the curse of dimensionality [19]. PCA addresses this issue by transforming the original variables into a smaller set of orthogonal components that preserve the dominant variance structure while improving interpretability [4,10,11].

For the clustering step, the relevant issue is that the number of meaningful regional groups is not known a priori. Affinity propagation was selected because it identifies exemplars directly from pairwise similarities and does not require a fixed cluster count in advance [5]. By contrast, k-means requires a pre-specified k and is sensitive to initialization [1], hierarchical clustering often yields subjective cut-off decisions and uneven group sizes [20], and DBSCAN is highly sensitive to parameter settings in small and structured datasets [7]. For the present study, the main value of affinity propagation lies in its ability to produce an interpretable regional partition without forcing a conventional k-based solution from the outset.

Taken together, these choices provide a transparent workflow for constructing a region-level spatial typology from correlated agricultural data. The contribution of the method lies less

in methodological novelty per se than in its ability to produce a geographically meaningful and empirically grounded classification for a large and internally heterogeneous country.

Validation. Validation in this study is primarily internal and comparative rather than predictive. First, cluster cohesion and separation were assessed in the main PC1–PC2 space using the Silhouette coefficient, the Davies–Bouldin index, and the Calinski–Harabasz index. Second, the three-cluster pattern was compared with alternative partitioning methods—k-means, Ward's hierarchical clustering, and PAM—to evaluate whether the broad regional structure was algorithm-specific or more generally stable. In this context, the preference value of −10 should be understood as an analyst-conditioned modeling choice aimed at obtaining a stable and geographically interpretable three-cluster solution, not as an automatic or uniquely determined parameter value.

Third, the resulting geography was interpreted against well-established contrasts in Kazakhstan's agricultural landscape, including the northern grain belt, lower-intensity dryland territories, and more diversified southern production systems. This strategy does not constitute external causal validation, and the results should therefore be interpreted as an empirically grounded regional typology rather than a predictive model. Its purpose is to test whether the classification is internally coherent, geographically plausible, and useful for regional comparison.

Results and its discussion. *Identification.* The first stage of the ISAC Analytical Process involves the systematic identification of the key variables to be used in the study, along with their corresponding data sources. This step aims to ensure the selection of indicators consistent with the conceptual framework of the research and to integrate heterogeneous data types into a coherent structure. In particular, the accurate definition of demographic, climatic, geographical, and production-related indicators is of critical importance for the classification of agricultural production regions. Therefore, the Identification stage should not be understood merely as a process of data collection, but rather as a strategic step that directly affects the reliability of the research and the robustness of subsequent analyses. In this section, some basic descriptive statistics of the variables are presented. The variables used in the analysis are grouped in Table 1.

Table 1

Variable groups used in the analysis (overview)

ID	Variables	Parameter Name
1	Demographic	Total Area
2		Population
3		Population Density
4	Agricultural	Agricultural Land Area
5		Number of Agricultural Enterprises
6		Agricultural Land per Enterprise
7		Annual Production Value
8		Annual Wheat Production
9	Climatic	Annual Average Temperature
10		Annual Precipitation
11		Annual Average Sunshine Duration
12	Geographic	Geographic Elevation

Note: The full data dictionary with precise definitions and units is provided in Appendix Table A2.

Table B1 (see Appendix B) presents the 12 demographic, climatic, geographic, and agricultural variables across the territorial units included for descriptive comparison. The wide ranges and substantial standard deviations across several indicators confirm that Kazakhstan exhibits pronounced internal heterogeneity rather than a uniform regional structure.

Segmentation. Principal Component Analysis (PCA) Results. Variance Explanation and Scree Plot. The PCA results demonstrate the explanatory power of the variables used in classifying Kazakhstan's agricultural production regions. The Scree Plot in Figure C1 (see Appendix C) shows that the first two components (PC1: 40%, PC2: 24%) account for 64% of the total variance, indicating that regional differences are largely captured by these components. While the inclusion of the third and fourth components increases the explained variance to nearly 80%, the contribution becomes marginal from the fifth component onward. The Elbow Criterion suggests that the additional information gained beyond the second component is limited, thereby confirming that PC1 and PC2 are sufficient for the dimensionality reduction process [9,11,14].

In summary, the results confirm that PCA reduces data dimensionality while retaining the main distinctions among agricultural regions, thereby providing a coherent basis for the subsequent biplot analysis.

Biplot Analysis of PCA. The biplot visualizes the first two principal components of PCA by jointly displaying observation scores and variable loading vectors, enabling interpretation of variable–region relationships [11,12]. In our results (Figure 1), PC1 loads strongly on agricultural land area, annual production value, wheat production, and the number of agricultural enterprises—capturing production capacity—whereas PC2 loads on average temperature, precipitation, and elevation—capturing environmental and geographic conditions.

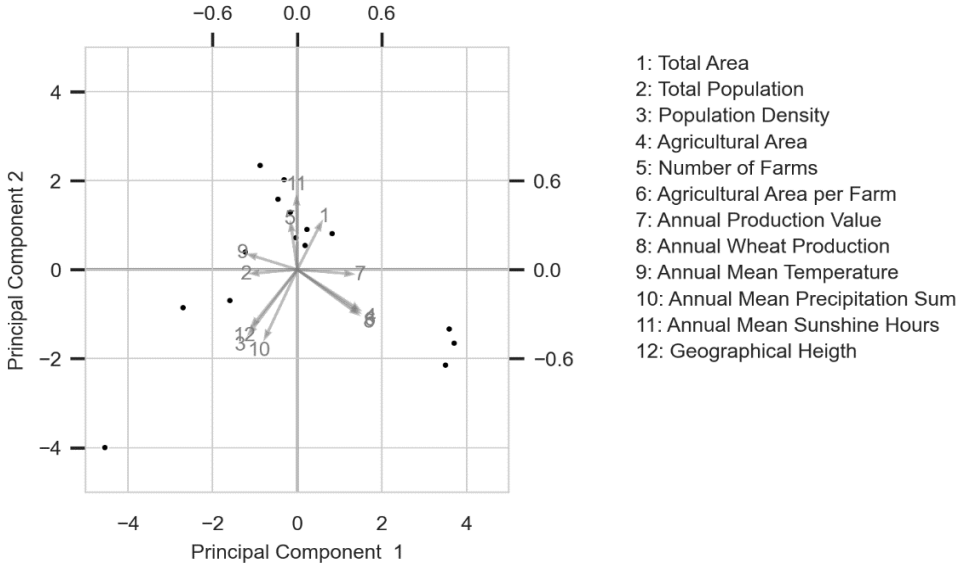


Figure 1. PCA biplot showing the first two principal components (64% cumulative variance)

Note: vectors represent variable loadings, and points indicate regional scores, revealing gradients in production capacity and climatic conditions.

In the second analysis, optimized by excluding certain variables (see Appendix C, Figure C2), only four agricultural variables ("annual production value," "annual wheat production," "agricultural land area," and "number of agricultural enterprises") were retained, with the first two components explaining 99% of the total variance. This exceptionally high proportion confirms the adequacy of the first two components for dimensionality reduction and the concentration of loadings in the Biplot.

The second biplot (Figure 2) more clearly illustrates the positions of these four variables within the PCA framework, thereby facilitating visual comparison of production patterns across regions.

The thematic maps (see Appendix C, Figure C3) depict the spatial distribution of these variables across Kazakhstan, revealing that the highest production values are concentrated in North Kazakhstan, Akmola, and Kostanay, while other regions exhibit diverse combinations.

Additionally, the spatial maps of all 12 variables (see Appendix C, Figure C4) visualize regional differences in agricultural production, climate, and topography.

The biplot analyses improve the comparative interpretation of regional differences by showing how production, climatic, and geographic indicators are distributed across the national territory. However, the main clustering solution is based on the optimized four-variable PCA representation, while the broader twelve-variable PCA is retained only for contextual interpretation and visualization.

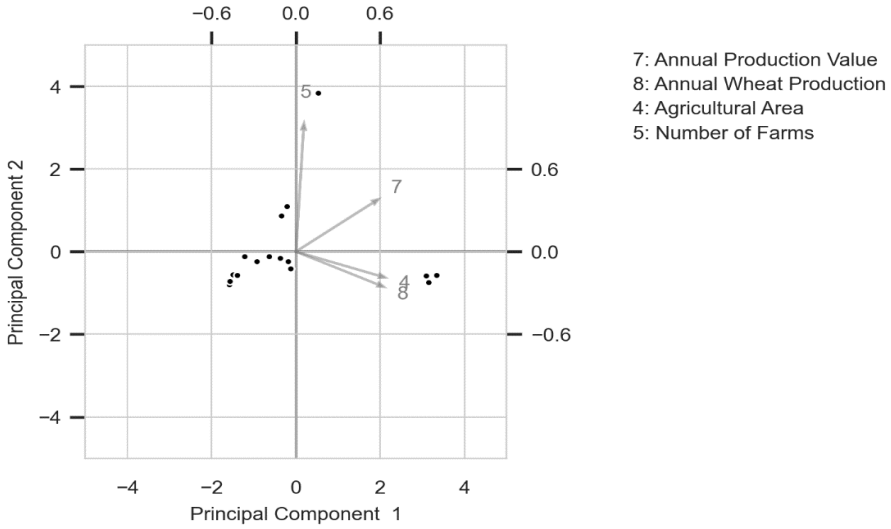


Figure 2. PCA Biplot of the four core agricultural variables

Note: arrows represent variable loading vectors, and points indicate regional scores on the first two components.

Cluster Analysis with Affinity Propagation (AP). The classification of agricultural production regions based on their shared characteristics is useful for regional comparison and for structuring differentiated planning discussions.

In this study, the Affinity Propagation (AP) clustering method was employed to classify Kazakhstan's agricultural production regions. AP performs clustering by considering pairwise similarities among data points and does not require the number of clusters to be specified in advance [5]. While AP does not require an explicit pre-specification of k, the preference parameter influences the number of exemplars and was tuned for stability and interpretability.

Table 2

AP settings and PC spaces used in clustering. Unless otherwise noted, we set similarity to −euclidean², set preference to a fixed value (preference = −10) to obtain a stable and substantively interpretable three-cluster solution, and used damping = 0.90; maximum iterations = 1000; convergence iterations = 50; random seed fixed for reproducibility

PCs used	Similarity	Preference	Damping	Max iter	Conv iter	Random seed	Resulting clusters (k)
PC1–PC2 (main)	−euclidean²	−10	0.90	1000	50	fixed	3 (consistent)
PC1–PC3 (sensitivity)	−euclidean²	median(s)	0.90	1000	50	fixed	—
PC1–PC4 (sensitivity)	−euclidean²	median(s)	0.90	1000	50	fixed	—

Note: PC1–PC2 is the main specification and yields three clusters (see Figure 3). Sensitivity runs with PC1–PC3/PC1–PC4 produced qualitatively similar cluster patterns and are omitted for brevity.

The first and second principal components (PC1 and PC2) derived from the optimized PCA results were used as inputs for the AP clustering analysis. Unless otherwise noted, AP clustering uses the PC1–PC2 scores from the four-variable PCA based on annual production value, annual wheat production, agricultural land area, and number of agricultural enterprises. This specification is intentionally focused on production structure and does not claim to represent the full agricultural-environmental complexity of Kazakhstan's regions. The broader twelve-variable PCA is reported separately for contextual interpretation and visualization.

Figure 3 shows cluster structure in the PC1–PC2 space; the spatial distribution of clusters is mapped in Figure 4.

Blue (Cluster 1) – Lower-intensity dryland-oriented regions.

Green (Cluster 2) – High-capacity grain regions.

Red (Cluster 3) – Distinct southern regions with a differentiated production profile.

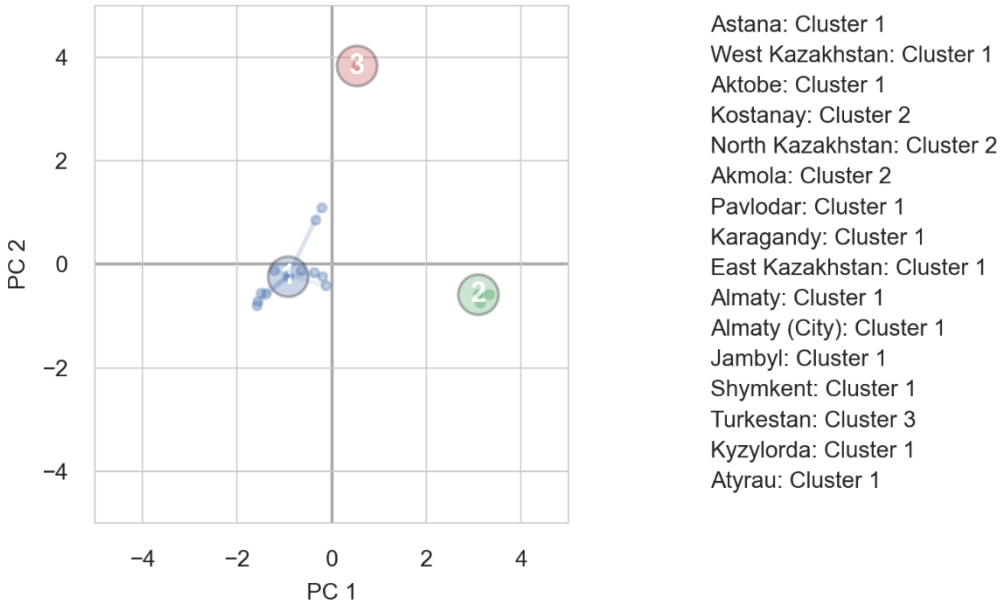


Figure 3. Affinity Propagation clusters of Kazakhstan's agricultural regions in the PC1–PC2 space. Blue = lower-intensity dryland-oriented regions; Green = high-capacity grain regions; Red = distinct southern regions with a differentiated production profile

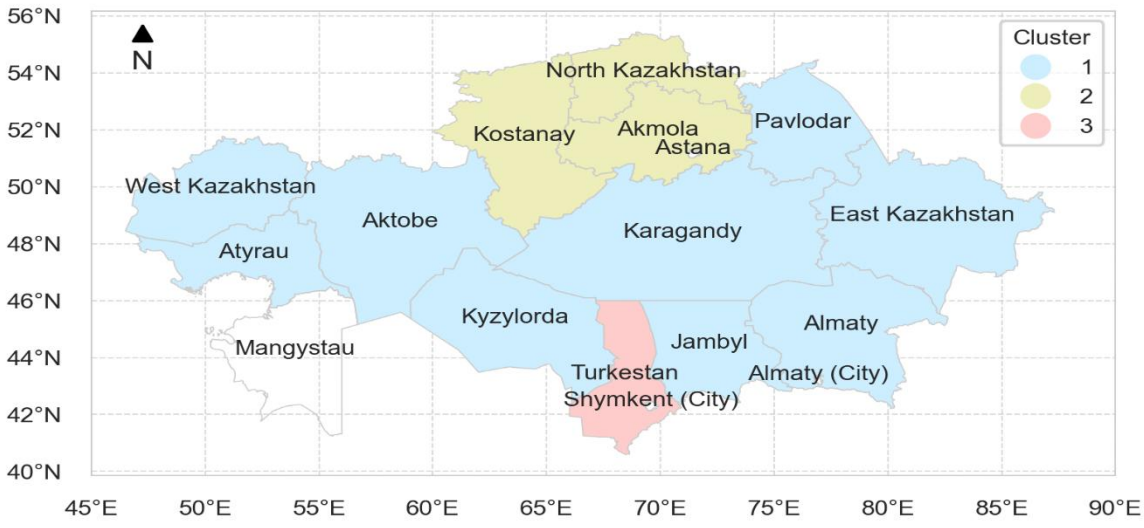


Figure 4. Spatial distribution of the three regional agricultural clusters in Kazakhstan (Blue = lower-intensity dryland-oriented regions; Green = high-capacity grain regions; Red = distinct southern regions with a differentiated production profile)

The variables used in the main clustering space—annual production value, annual wheat production, agricultural land area, and number of agricultural enterprises—capture key differences in regional production structure. Interpreted together with the broader contextual PCA, the AP clustering results help clarify how Kazakhstan's agricultural regions differ in terms of production intensity and wider territorial conditions.

We assessed cluster quality using Silhouette, Davies–Bouldin, and Calinski–Harabasz indices; results indicate coherent separation across the three clusters (Silhouette = 0.723; Davies–Bouldin = 0.192; Calinski–Harabasz = 47.64). As an additional qualitative check, we compared the overall three-cluster pattern against k-means (k=3), Ward's hierarchical clustering (k=3), and PAM (k=3) in the PC1–PC2 space and found that the broad cluster structure remains comparable, although point-wise assignments may vary by algorithm.

Characterization of Regional Agricultural Clusters. The three clusters comprise 16 territorial units and reflect differences in production structure and broader territorial conditions. The AP clustering results indicate that Kazakhstan's agricultural production geography can be grouped into three major regional configurations. These configurations should be understood primarily as production-oriented clusters that are subsequently interpreted in relation to their wider territorial context.

Cluster 1 (Blue): Lower-intensity dryland-oriented regions

This cluster comprises regions such as Atyrau, West Kazakhstan, Karaganda, East Kazakhstan, Kyzylorda, and Pavlodar, which occupy the lower-intensity side of the production-oriented clustering space. In comparative terms, these regions are less dominated by the concentrated grain-production structure that defines the northern cluster and are more closely associated with broader territorial conditions commonly linked to dryland agricultural settings. The inclusion of East Kazakhstan in this cluster should be interpreted in relation to the multivariate production structure used in the clustering model rather than to single-sector expectations based on localized grain-producing areas within the region.

Cluster 2 (Green): High-capacity grain regions

This cluster encompasses Kostanay, Akmola, and North Kazakhstan, which stand out clearly in the optimized production-space clustering and correspond to the country's highest-output grain-producing territories.

The regions in this cluster are distinguished by high wheat production, large agricultural land area, high production value, and a large-enterprise structure in the optimized PCA space. In comparative terms, they represent the most concentrated grain-production context in the national typology. Their classification is therefore best understood as a high-capacity grain-oriented regional grouping rather than as a direct statement about specific soil conditions, management practices, or environmental outcomes.

Cluster 3 (Red): Distinct southern regions with a differentiated production profile

This cluster captures a distinct southern configuration centered on Turkistan and nearby territories. In the clustering results, it is separated from both the northern grain cluster and the broader lower-intensity group, indicating a southern production structure that does not fit the logic of the other two regional configurations.

Given the available variables, this cluster should not be overinterpreted in sectoral terms. What can be stated with confidence is that it represents a territorially distinct southern grouping that does not conform either to the concentrated northern grain structure or to the broader lower-intensity dryland-oriented cluster. Its main analytical value lies in signaling that southern regions should not automatically be discussed within the same planning logic as the other two cluster types.

Figure 4 shows that the typology reproduces broad regional contrasts that are already familiar in Kazakhstan's agricultural geography, including the concentration of grain production in the north and the differentiated production profile of southern regions. The added value of the PCA–AP workflow is not that it discovers an entirely unknown geography, but that it renders these contrasts operational within a transparent multivariate framework. In practical

terms, this makes it possible to compare non-contiguous regions on the basis of shared structural characteristics rather than administrative proximity alone. From a regional geographic analysis perspective, the contribution of the model therefore lies in converting territorially uneven agricultural conditions into a reproducible classification that can support regional comparison and differentiated planning.

A compact characterization of the three clusters is provided in Table C1, including qualitative descriptors for their main structural features.

The clustering results show that Kazakhstan's agricultural regions diverge along production and broader territorial gradients. The identified groups distinguish the highly productive northern grain regions, lower-intensity dryland-oriented territories, and a distinct southern production structure.

Rather than serving as a direct policy prescription, the typology provides a spatial starting point for more differentiated regional planning discussions, including comparative assessment of grain-system territories, infrastructure prioritization, and the targeting of future field-based research.

Why three clusters? Given the limited number of territorial units and the compact indicator structure used in the main specification, affinity propagation converged to a three-cluster solution that was both stable and geographically interpretable. This reflects a trade-off between parsimony and differentiation at the regional scale. More fine-grained partitions would likely emerge with smaller territorial units or a broader socio-economic indicator set.

Discussion in Relation to Agricultural Regionalization. The findings contribute to the literature on agricultural regionalization by showing how a parsimonious multivariate workflow can be used to construct a geographically interpretable regional typology at the national scale. Unlike expert-based agro-ecological frameworks, which primarily assess biophysical suitability, the present typology is based on observed regional characteristics and is therefore more directly aligned with questions of territorial differentiation in agricultural structure. Its purpose is not to replace AEZ-type frameworks, but to complement them with a region-level classification that is more sensitive to within-country variation.

From the perspective of regional geographic analysis, the main contribution of the study lies in its treatment of agricultural planning as a spatial problem of comparison, allocation, and differentiation. The results show that regions that are geographically distant can still share structurally similar agricultural profiles, while neighboring units may belong to different functional groupings. This matters because planning based solely on administrative or contiguous territorial logic risks obscuring differences in production intensity and broader territorial context.

The three-cluster solution should nevertheless be interpreted with caution. It is coherent and geographically plausible, but it is not the only conceivable classification of Kazakhstan's agricultural territory. It is best understood as a baseline regional typology derived from the available data and the chosen level of aggregation. Finer-grained territorial units, additional socio-economic variables, or alternative model specifications would likely produce more differentiated partitions. The present contribution therefore lies in building a transparent starting point for regionally differentiated analysis rather than in claiming a definitive agricultural map of the country.

Implications for Regional Resource Planning. The practical relevance of the typology lies in its ability to support differentiated regional planning rather than one-size-fits-all interpretations of Kazakhstan's agricultural territory. The three-cluster structure suggests that northern grain-producing regions, broader lower-intensity dryland-oriented territories, and the distinct southern grouping should not automatically be treated as equivalent planning contexts. In this sense, the classification helps narrow the comparison space by indicating which territories are structurally similar enough to be discussed together and which are not.

These implications should not be overstated as direct policy prescriptions. The typology does not identify optimal interventions by itself. Its value lies in narrowing the comparison

space and in providing a spatial baseline for more targeted regional analysis, including comparative assessment of grain-system territories, infrastructure prioritization, and the targeting of future field-based research.

The broader analytical relevance of the study is correspondingly modest but clear. It shows how regional geographic reasoning can be operationalized through a transparent statistical workflow that links territorial heterogeneity to planning relevance.

Study Limitations and Future Research Directions. This study should be interpreted as a region-level classification exercise, not as a causal or predictive model of agricultural performance. Several limitations follow from that design choice. First, the clustering is based on a relatively small number of territorial units, which constrains the degree of differentiation that can be identified and limits statistical generalization. Second, although the empirical basis spans 2015–2023, the typology is constructed from aggregated regional profiles and therefore does not capture temporal dynamics directly. Third, the main clustering solution relies on a reduced production-structure variable set, which improves parsimony but also narrows the meaning of the resulting classification.

Additional limitations concern the thematic scope of the available data. The present typology does not directly incorporate market access, subsidy structures, irrigation intensity, farm income, land degradation, or remote-sensing indicators of land-use change. As a result, the classification should not be read as a complete account of Kazakhstan's agricultural geography, but rather as a structured baseline derived from the information that was consistently available at the national scale.

For this reason, the present typology is best treated as a starting point for further geographical analysis. Future work could extend it by incorporating finer spatial units, remote-sensing indicators, market-access variables, irrigation intensity, and land-degradation proxies. Such extensions would be especially valuable for testing how stable the observed clusters remain when spatial resolution and explanatory dimensions are expanded.

Conclusions. This study developed a data-driven spatial typology of agricultural regions in Kazakhstan using principal component analysis and affinity propagation clustering. Based on region-level data for 2015–2023, the analysis identified three interpretable clusters that distinguish highly productive northern grain regions, lower-intensity dryland-oriented territories, and a distinct southern production structure.

The contribution of the article is practical and geographical rather than methodological. It provides a transparent framework for classifying subnational agricultural heterogeneity and for supporting regional comparison beyond national averages alone. Internal validation metrics indicated coherent cluster separation in the main specification (Silhouette = 0.723; Davies–Bouldin = 0.192; Calinski–Harabasz = 47.64), supporting the interpretability of the three-cluster solution.

At the same time, the classification should be read as a baseline typology rather than a definitive territorial model. It is based on aggregated regional profiles and a production-oriented clustering space, and it does not capture temporal dynamics or the full environmental and socio-economic complexity of Kazakhstan's agricultural geography. Even so, it provides a usable starting point for more differentiated regional analysis and for future refinement through finer spatial units and broader indicator sets.

References:

1. Alam, M., & Perumal, T. (2021), Performance analysis of affinity propagation clustering for IoT-based big data. *IEEE Access*, 9, 15062–15077. https://doi.org/10.1109/ACCESS.2021.3052651

2. Altieri, M. A.; Nicholls, C. I. (2017), The adaptation and mitigation potential of traditional agriculture in a changing climate. *Clim. Change*, *140*(1), 33–45. https://doi.org/10.1007/s10584-013-0909-y

3. Baisholanov, S. Y.; Kokhan, S.; Shakhanov, E.; Scherbinin, A.; Zhakashov, A. (2024), Agro-Climatic Zoning of the Territory of Northern Kazakhstan and Development of an Assessment System for the Agro-Climatic Resources of Dryland Crop Production. *Climate*, *12*(8), 146.

4. Ding, C.; He, X. (2004), K-means clustering via principal component analysis. In *Proceedings of the 21st International Conference on Machine Learning (ICML 2004)*; p 29. https://doi.org/10.1145/1015330.1015408

5. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. (1996), A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise (DBSCAN). In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*; AAAI Press: Portland, OR, pp 226–231.

6. Frey, B. J.; Dueck, D. (2007), Clustering by passing messages between data points. *Science, 315*(5814), 972-976. https://doi.org/10.1126/science.1136800

7. Ghosh, S.; Halder, S.; Das, S. (2022), Affinity propagation clustering for complex networks: A comparative study, *Appl. Intell.,* 52(8), 8462–8484. https://doi.org/10.1007/s10489-021-02659-4

8. Ghosh, S.; Hasan, S. S.; Rahman, M. M. (2021), Regional heterogeneity in agricultural productivity and food security: Evidence from South Asia. *Sustainability*, *13*(21), 11754. https://doi.org/10.3390/su132111754

9. Hair, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E. (2019), *Multivariate Data Analysis*, 8th ed.; Cengage Learning: Boston, MA.

10. IIASA; Food and Agriculture Organization of the United Nations. (2021), *GAEZ—Global Agro-Ecological Zones v4. Documentation and Data Portal;* Available online: https://openknowledge.fao.org/items/039f7ec9-98af-49e1-8d24-850122c69bef (accessed on 20 February 2026).

11. Jolliffe, I. T. (2002), *Principal Component Analysis*; Springer Series in Statistics; Springer: New York. https://doi.org/10.1007/b98835

12. Jolliffe, I. T., Cadima, J. (2016), Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc., 374* (2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

13. Lorbieski, R.; Guedes, L. P. C.; Uribe-Opazo, M. A.; Kestring, F. B. F. (2023), Regionalization of an Agricultural Area by Means of Multivariate Data and Their Relationship with Soybean Productivity. *Aust. J. Crop Sci.*, *17* (7), 570–580. https://doi.org/10.21475/ajcs.23.17.07.p3895

14. Rahman, A. S.; Rahman, A. (2020), Application of Principal Component Analysis and Cluster Analysis in Regional Flood Frequency Analysis: A Case Study in New South Wales, Australia. *Water*, *12* (3), 781. https://doi.org/10.3390/w12030781

15. Savin, I.; Kobets, A.; Kuzmin, A. (2025), Organic Crop Production in Kazakhstan: Current State and Prospects. *Resources*, *14* (1), 10. https://doi.org/10.3390/resources14070108

16. Shukla, A.K.; Behera, S. K.; Basumatary, A.; Sarangthem, I.; Mishra, R.; Dutta, S.; Sikaniya, Y.; Sikarwar, A.; Shukla, V.; Datta, S. P. (2024), PCA and Fuzzy Clustering-Based Delineation of Soil Nutrient Management Zones. *J. Environ. Manage. 361*, 120982. https://doi.org/10.1016/j.jenvman.2024.121511

17. Verleysen, M.; François, D. (2005), The Curse of Dimensionality in Data Mining and Time Series Prediction. In *Proceedings of the International Work-Conference on Artificial Neural Networks*; Springer: Berlin, Heidelberg; pp 758–770. https://doi.org/10.1007/11494669_93

18. Ward, J. H. (1963), Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc. 58* (301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

19. Zhang, J.; Wang, X.; Li, J. (2020) An Improved Affinity Propagation Clustering Algorithm for Mixed Data. *Knowl.-Based Syst.*, *195*, 105665. https://doi.org/10.1016/j.knosys.2020.105665

Appendix A

Appendix A.1 Data Dictionary

Various software tools were employed to process, model, and visualize the different types of data used in the analysis. The Python programming language served as the primary tool for data preprocessing, statistical modeling, and dimensionality reduction, with particular reliance on the NumPy, pandas, and scikit-learn libraries. During the dimen-sionality reduction stage, Principal Component Analysis (PCA) was applied to decrease multicollinearity among variables, minimize information loss, and render the dataset more suitable for clustering

algorithms. For the clustering step, Affinity Propagation (AP) was employed, enabling the grouping of agricultural production regions based on similarities without the need to predetermine the number of clusters. QGIS software was used for processing, mapping, and spatial analysis of location-based data, while Matplotlib was applied to present the findings graphically and support the interpretation of results.

Table A2

Indicators used in the analysis (Data Dictionary: definitions, units, transformations, sources).

0	Indicator	Definition	Unit	Transform/ Scale	Temporal agg.	Spatial agg.	Source	Notes
1	Total Area	Total land area of each administrative unit (region/city).	km^2	winsor p1–p99; z-score	static (2023)	region/city	KazStat; FAO	Static variable; used for density calculations and context.
2	Population	Total resident population of the administrative unit.	persons	winsor p1–p99; z-score	annual mean 2015–2023	region/city	KazStat	If only annual snapshots exist, average across 2015–2023.
3	Population Density	Population per square kilometer.	persons/km^2	winsor p1–p99; z-score	annual mean 2015–2023	region/city	KazStat	Derived as Population / Total Area when not provided.
4	Agricultural Land Area	Total agricultural land area (cropland, pasture as defined by source).	ha	winsor p1–p99; z-score	annual mean 2015–2023	region/city	KazStat; FAO	Confirm land-use definition consistency across years.
5	Number of Agricultural Enterprises	Number of registered agricultural enterprises.	count	winsor p1–p99; z-score	annual mean 2015–2023	region/city	KazStat	Count-based indicator; ensure consistent registry definition.
6	Agricultural Land per Enterprise	Average agricultural land area per enterprise.	ha/enterprise	winsor p1–p99; z-score	annual mean 2015–2023	region/city	KazStat	Computed as V4 / V5 where necessary; check division by zero.
7	Annual Production Value	Monetary value of total agricultural production per year.	million KZT	winsor p1–p99; z-score	annual mean 2015–2023	region/city	KazStat; OECD	If needed, adjust for inflation to constant prices (note year).
8	Annual Wheat Production	Total wheat produced annually.	10^3 t	winsor p1–p99; z-score	annual mean 2015–2023	region/city	KazStat; FAO	Converted from tons to 10^3 t to avoid large magnitudes.
9	Annual Average Temperature	Annual mean near-surface air temperature.	°C	winsor p1–p99; z-score	annual mean 2015–2023	region/city	Kazhydromet; NASA	Ensure same grid/source across years; area-weight as needed.

10	Annual Precipitation	Total precipitation accumulated over the year.	mm/year	winsor p1–p99; z-score	annual mean 2015–2023	region/city	Kazhydromet; NASA	If multiple sources, reconcile and document method.
11	Mean Daily Sunshine Duration	Average daily sunshine duration.	h/day	winsor p1–p99; z-score	annual mean 2015–2023	region/city	Kazhydromet	Standardized from hours/year → h/day for interpretability.
12	Geographic Elevation	Area-weighted mean elevation of the administrative unit.	m	winsor p1–p99; z-score	static (SRTM epoch)	region/city	NASA SRTM; QGIS Analysis	Static terrain metric; document resampling and masks.

Notes on units: wheat production is reported in 10^3 t; precipitation in mm/year; sunshine in h/day (mean daily); production value in million KZT (specify if constant prices and the base year). All variables are winsorized at p1–p99 and z-scored prior to PCA unless otherwise stated.

Appendix B. Region-Level Statistics

Table B1 Selected indicators by territorial unit.

(Note: The three republican cities and the post-2022 regions Abai, Ulytau, and Zhetysu are shown for descriptive context only. Clustering analyses were conducted on the 16-unit analytical sample defined in Section 2, which excludes Mangystau and uses the stable pre-2022 territorial configuration.)

Regions	***Demographic***			***Agricultural***				
	1	2	3	4	5	6	7	8
Astana	720	1,184,469	1,486	1,799	1,548	1	509	1
West Kazakhstan	151,300	661,172	4	586,158	9,908	59	246,477	130
Aktobe	300,629	893,669	3	717,900	9,897	73	379,902	190
Kostanay	196,001	864,529	4	5,256,509	7,859	669	879,105	4,340
North Kazakhstan	97,993	543,679	6	4,406,890	6,463	682	1,018,254	4,365
Akmola	146,219	735,481	5	5,397,297	9,591	563	973,591	4,300
Pavlodar	124,755	751,011	6	1,568,885	6,803	231	396,862	786
Karaganda	427,982	1,375,788	3	1,192,528	9,626	124	415,709	850
Ulytau (ex Karaganda)				39,426	3,488	11	77,077	
East Kazakhstan	283,226	1,363,656	5	641,816	9,605	67	427,834	790
Abai (ex East Kazakhstan)				742,558	10,386	71	371,258	
Almaty	223,911	2,077,656	9	517,224	29,997	17	648,899	215
Zhetysu (ex Almaty)				454,224	19,984	23	425,524	
Almaty City	682	1,977,011	2,898	627	3,069	0	8,154	0
Zhambyl	144,264	1,139,151	8	756,039	27,986	27	526,962	195
Shymkent	1,163	1,074,167	923	23,494	5,924	4	39,454	8
Turkestan	116,095	2,044,511	18	859,907	83,749	10	945,882	219
Kyzylorda	226,019	814,461	4	188,545	13,141	14	161,312	11
Atyrau	118,631	657,118	6	12,097	3,901	3	110,627	0
Mean	159,974	1,134,846	337	1,229,680	14,364	139	423,863	1,025
Std Deviation	112,287	492,758	776	1,701,094	17,996	223	326,111	1,615
Min	682	543,679	3	627	1,548	0	509	0
Max	427,982	2,077,656	2,898	5,397,297	83,749	682	1,018,254	4,365

Regions	*Climatic*			*Geographic*
	9	10	11	12
Astana	3	307	6	338
West Kazakhstan	6	348	6	35
Aktobe	4	311	7	225
Kostanay	4	336	7	155
North Kazakhstan	2	353	6	140
Akmola	3	314	2	234
Pavlodar	3	313	10	123
Karaganda	4	325	7	546
Ulytau (ex Karaganda)	-	-	-	674
East Kazakhstan	4	433	12	298
Abai (ex East Kazakhstan)	14	329	11	67
Almaty	8	370	6	600
Zhetysu (ex Almaty)	-	-	-	675
Almaty City	10	674	2	1,100
Zhambyl	11	355	12	623
Shymkent	13	576	8	506
Turkestan	13	203	9	212
Kyzylorda	8	150	8	128
Atyrau	10	177	7	-20
Mean	7	346	7	350
Std Deviation	4	124	3	283
Min	2	150	2	-20
Max	14	674	12	1,100

Appendix C. Supplementary figures and cluster summary

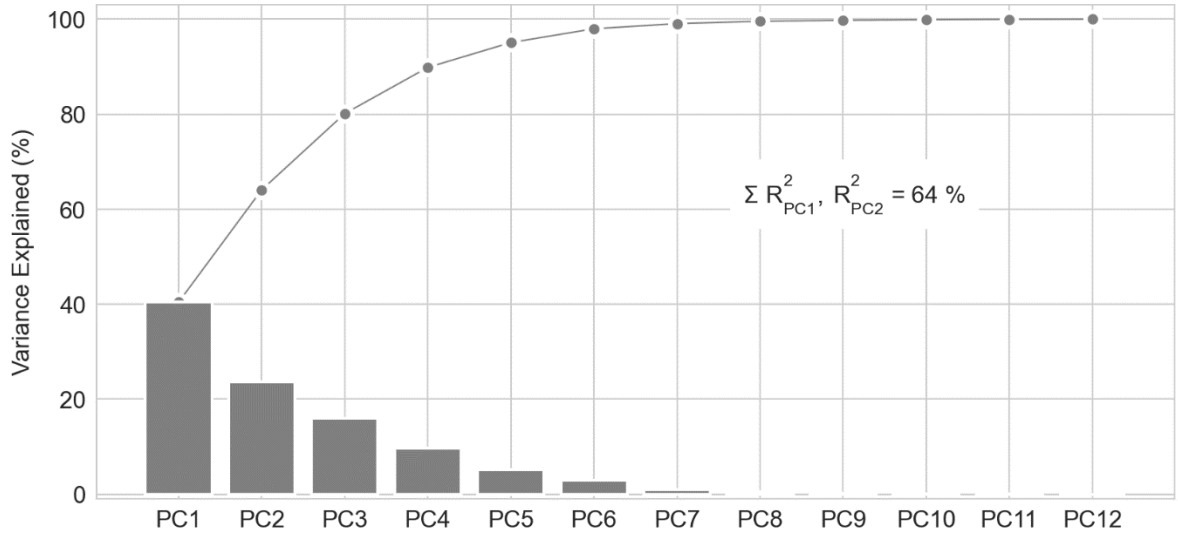


Figure C1
Scree Plot illustrating the explained variance of principal components.

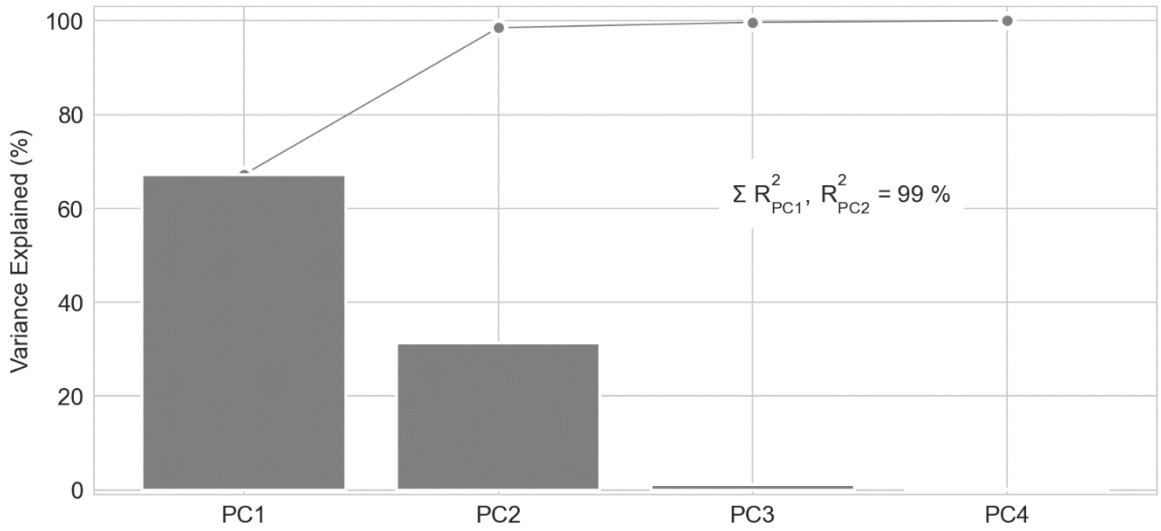


Figure C2. Scree Plot of PCA with the optimized variable set (annual production value, wheat production, agricultural land area, and number of agricultural enterprises); the first two components explain 99% of the total variance.

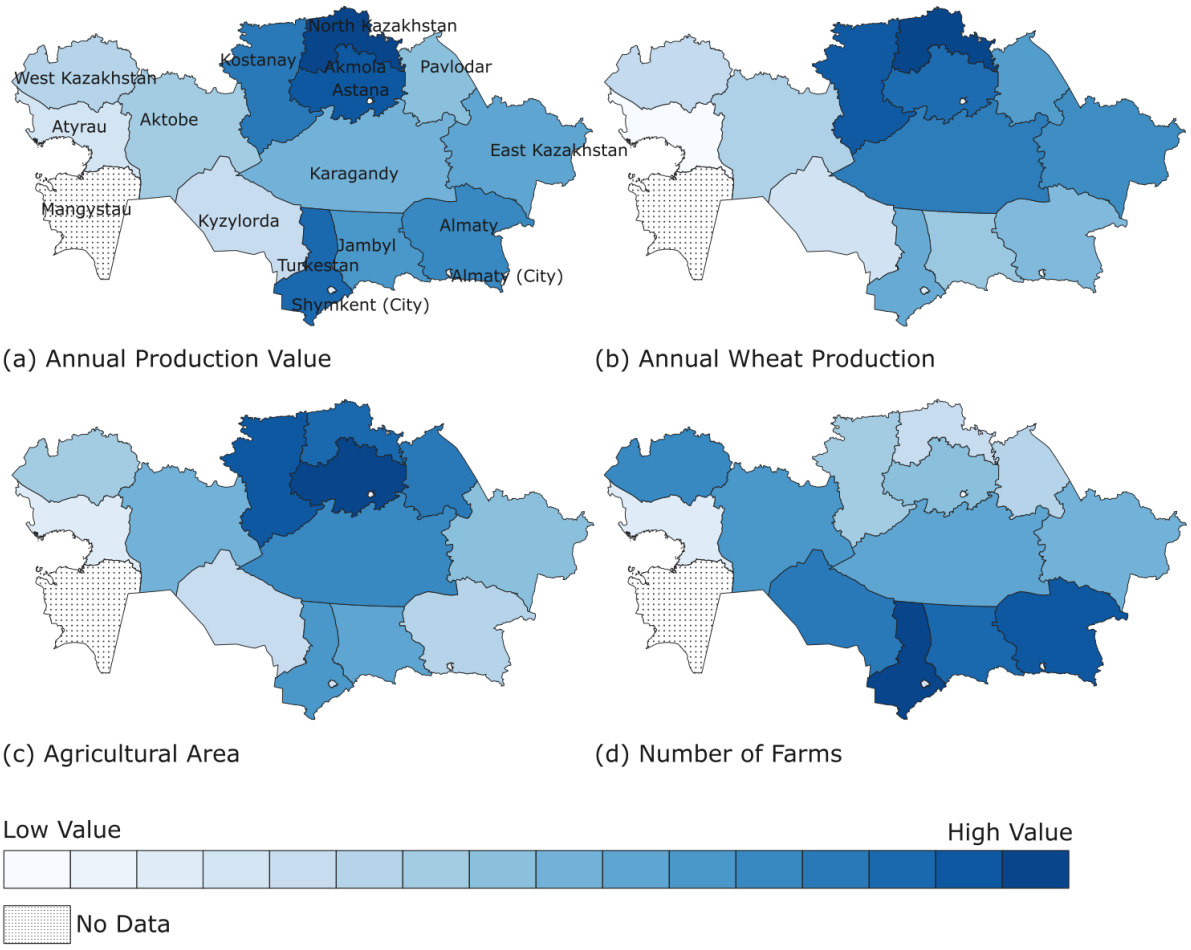


Figure C3. Spatial distribution of the four core agricultural variables.

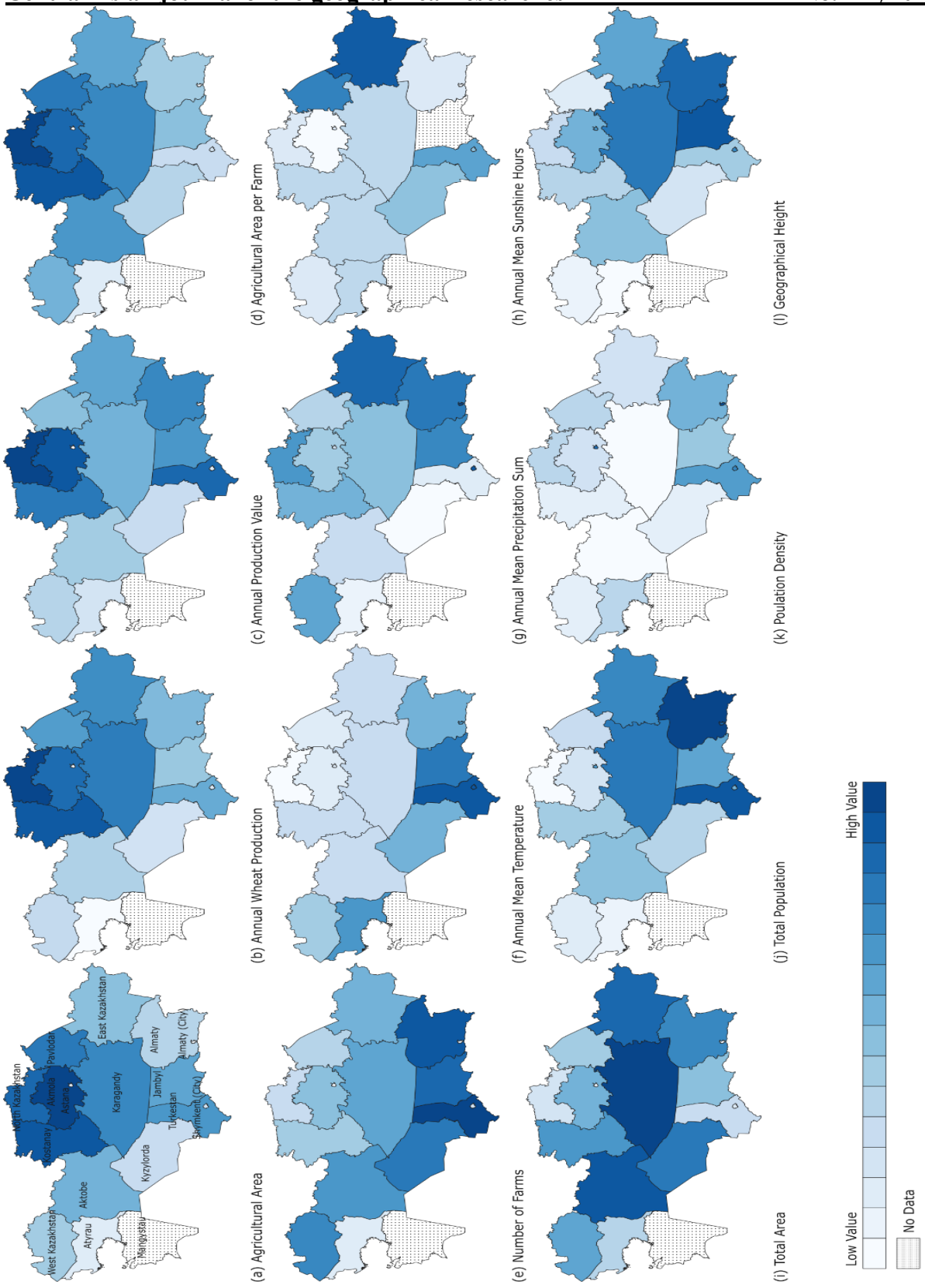


Figure C4. Regional-level characteristics of agricultural, demographic, climatic, and topographic variables in Kazakhstan. The spatial representation of these 12 variables was examined through principal component analysis.

Table C1

Cluster characterization at a glance (descriptive summary).

Cluster	Short label	Member units	Dominant features	Interpretive note
Cluster 1	Lower-intensity dryland-oriented regions	see Figure 4	Lower production value and wheat output; broader dryland-oriented territorial context	Best interpreted as a lower-intensity regional grouping rather than a single uniform production system
Cluster 2	High-capacity grain regions		Higher wheat production, larger agricultural land area, higher production value	Represents the most concentrated grain-production structure in the typology
Cluster 3	Distinct southern regions with a differentiated production profile		Separated from both the northern grain cluster and the broader lower-intensity grouping	Signals a territorially distinct southern context requiring separate comparison

Information about the authors:

Sami Celtikoglu — Independent Institute for Environmental Issues (UfU e.V.) (Berlin, Germany); Leibniz Institute of Agricultural Development in Transition Economies (IAMO) (Halle (Saale), Germany); Humboldt-Universität zu Berlin (Berlin, Germany), External Doctoral Researcher. E-mail: sami.celtikoglu@ufu.de

Murat Cankurt — North Carolina Agricultural and Technical State University (North Carolina, USA), PhD, Associate Profeccor. E-mail: mcankurt@ncat.edu

Arne Reck — Leibniz Institute of Agricultural Development in Transition Economies (IAMO) (Halle (Saale), Germany), PhD. E-mail: arne.reck@ufu.de

Yessengali Oskenbayev — Narxoz University (Almaty, Kazakhstan), PhD, Associate Professor. E-mail: yessengali.oskenbayev@narxoz.kz

Terrence Thomas — North Carolina Agricultural and Technical State University (North Carolina, USA), Dr, Professor. E-mail: twthomas@ncat.edu

Сведения об авторах:

Челтикоглу Сами — Независимый институт экологических проблем (UfU e.V.) (Берлин, Германия); Институт сельскохозяйственного развития в странах с переходной экономикой имени Лейбница (IAMO) (Халле (Заале), Германия); Берлинский университет имени Гумбольдта (Берлин, Германия), внешний докторант. E-mail: sami.celtikoglu@ufu.de

Чанкурт Мурат — Сельскохозяйственный и технический государственный университет Северной Каролины (Северная Каролина, США), доктор философии (PhD), ассоциированный профессор. E-mail: mcankurt@ncat.edu

Рек Арне — Институт сельскохозяйственного развития в странах с переходной экономикой имени Лейбница (IAMO) (Халле (Заале), Германия), доктор философии (PhD). E-mail: arne.reck@ufu.de

Оскенбаев Ессенгали — Университет Нархоз (Алматы, Казахстан), кандидат экономических наук, ассоциированный профессор. E-mail: yessengali.oskenbayev@narxoz.kz

Томас Терренс — Северо-Каролинский сельскохозяйственный и технический государственный университет (Северная Каролина, США), доктор, профессор. E-mail: twthomas@ncat.edu

For citation:

Celtikoglu S., Cankurt M., Reck A., Oskenbayev Y., Thomas T. (2026), A data-driven spatial typology of agricultural regions in Kazakhstan: a pca and affinity propagation approach, *Central Asian Journal of Geographical Researches,* No. 1-2, pp. 40-58.

Для цитирования:

Челтикоглу С., Джанкурт М., Рек А., Оскенбаев Е., Томас Т. Пространственная типология сельскохозяйственных регионов Казахстана, основанная на данных: подход с использованием метода главных компонент и распространения сходства // Центральноазиатский журнал географических исследований. 2026. № 1-2. С. 40-58. (На англ. яз.).