

Hierarchical properties and mechanisms affecting quality evaluation outcomes of reconstructed prehistoric roads under incomplete data

Bo Wu

wubo@hzu.edu.cn

Huizhou University

Qiong Zhang

Huizhou University

Yitong Hong

Huizhou University

Haojia Gao

Huizhou University

Zhihua Zhu

Huizhou University

Hui Yin

Huizhou University

Dongning Feng

Huizhou University

Jiashuo Chen

Huizhou University

Ruixi Zhu

Huizhou University

Ziqiao Kang

Huizhou University

Article

Keywords: incomplete archaeological data, evaluation mechanisms, training and test set sizes, settlement spatial patterns, hierarchical length scaling, prehistoric road networks

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10037844/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

**Hierarchical properties and mechanisms affecting quality
evaluation outcomes of reconstructed prehistoric roads under
incomplete data**

**Bo Wu^{1*}, Qiong Zhang¹, Yitong Hong¹, Haojia Gao¹, Zhihua Zhu¹, Hui Yin¹,
Dongning Feng¹, Jiashuo Chen¹, Ruixi Zhu¹, Ziqiao Kang¹**

¹School of Geography and Tourism, Huizhou University, Huizhou 516007, China

*Corresponding author: Bo Wu (wubo@hzu.edu.cn)

Abstract

Reconstructing prehistoric road networks is essential for understanding ancient spatial organization, yet evaluation is constrained by incomplete settlement data. Using Neolithic site records from the Haidai and circum-Taihu regions, we partition observations into training and independent test sets and assess reconstructions via a settlement interaction model. Controlled experiments show that, once training data are sufficient, accuracy approaches saturation, whereas reliability scales with the square root of test-set size, supporting a conservative interpretation of evaluation outcomes under incomplete data. Cross-regional comparisons further indicate that settlement clustering and spatial unevenness modulate this relationship. Applying the framework to five Neolithic periods in Haidai, we find that exponential hierarchical length scaling is statistically ambiguous in the Houli period, clearly favored in Beixin, and becomes progressively clearer from Dawenkou onward, suggesting that road-network hierarchy emerged gradually as settlement systems became denser and more complex.

Keywords: incomplete archaeological data; evaluation mechanisms; training and test set sizes; settlement spatial patterns; hierarchical length scaling; prehistoric road networks

Introduction

The formation and evolution of ancient transportation networks are fundamental to understanding how societies organize space. Roads connect places, and they are deeply implicated in resource allocation, population mobility, technological transmission, and political integration^[1].

Consequently, their spatial structure records, to a considerable extent, the modes of organization and evolutionary trajectories of societies in different regions^[2–5]. For the Neolithic and Bronze Ages, which lack continuous written records and unambiguous road-surface remains, ancient road networks are seldom amenable to direct reconstruction through traditional archaeological excavation. It is therefore necessary to rely on geographic information systems, spatial interaction models, and complex network analysis to infer potential corridors and their network configurations indirectly from known settlement distributions, environmental constraints, and inter-site connections^[6,7].

The methodological framework of this field has progressively expanded from point-to-point route inference to the modeling and interpretation of regional networks. Earlier work was largely grounded in cost surfaces and least-cost path analysis, translating slope, surface roughness, drainage density, vegetation cover, and other impediments to movement into mobility costs, and on this basis identifying potential corridors between two or more locations^[8–13]. This approach has been widely applied to issues such as ancient road identification, regional accessibility analysis, trade route discussion, and the study of early dispersal processes, including Roman roads, pre-Columbian pedestrian transportation networks, the Silk Road in the Asian highlands, and early hominin routes out of Africa^[14–16]. At the same time, researchers have recognized that the

specification of the cost function is not a purely technical step, but directly affects the interpretation of ancient mobility behavior. Walking time, energy expenditure, and fatigue burden are not equivalent, and the behavioral assumptions corresponding to different indicators are likewise distinct; if the physiological constraints and decision-making logic of movers are ignored, the interpretive power of path simulation is correspondingly limited^[17,18]. Moreover, the formation of historical roads does not always follow a simple least-effort principle in topographic terms. Their spatial trajectories are often jointly shaped by fords, nodal settlements, institutionalized corridors, and the continuity of pre-existing paths^[12,19].

As the scale of research expanded from single paths to regional networks, the interpretive limitations of traditional least-cost path methods became increasingly apparent, particularly their reliance on predefined origins and destinations and on single-path output. To overcome this constraint, researchers introduced multiple-path and whole-network generation methods to characterize more continuous and more distributed mobility potential within regions. Howey^[20] demonstrated that circuit theory can complement the multipath movement possibilities that least-cost path analysis cannot readily express. The FETE model proposed by White and Barber^[14] attempted to generate movement networks closer to natural formation processes without prespecifying explicit origins and destinations. By simulating the seasonal mobility of pastoral groups, Frachetti et al.^[15] revealed the formation mechanisms of ecologically driven mobility networks in highland regions. Correspondingly, spatial interaction models have been introduced into archaeological research with increasing frequency. Their focus is no longer on identifying the lowest-cost path, but rather on understanding how the strength of connections between places is formed and maintained under geographical constraints.

Related studies emphasize that socio-spatial networks are not arbitrary collections of connections; their generation and maintenance always involve explicit or implicit cost allocation, and network modeling must therefore address simultaneously the coupled relationships among node attributes, spatial resistance, and interaction mechanisms^[6]. Recent work on road network reconstruction in the Haidai region has further combined settlement interaction models, least-cost paths, and flow-accumulation methods to simulate the formation of road systems from the Bronze Age to the Early Iron Age at a relatively large spatiotemporal scale, and has discussed their relationship with population growth, settlement expansion, and the intensification of regional interaction^[21].

The expansion of scale from paths to networks has also shifted the research focus from path identification to network-structure interpretation. Complex network theory indicates that the growth of spatial networks is not a simple accumulation of edges, but may be accompanied by systematic changes in connectivity patterns, hierarchical structure, and organizational efficiency^[22,23]. In archaeology, complex network approaches have been used to analyze prehistoric exchange relations, raw-material circulation, and patterns of regional interaction, and to reveal differences among network types in terms of small-world properties, centrality, and community structure^[3,24]. This perspective suggests that road networks are not merely geometric expressions of mobility behavior, but may also constitute spatial manifestations of the evolution of regional societies from loose connections toward more highly organized structures. General studies of spatial networks show that, in mature transport systems, road length, quantity, and grade often exhibit relatively stable hierarchical relationships, and their growth processes may also display pronounced scaling characteristics^[25,26]. The relationship between road grade and length

therefore reflects not only morphology, but also potentially the developmental stage of the network and the level of social organization underlying it.

However, the validity of discussions about network developmental characteristics depends on whether the reconstructed network has undergone rigorous validation. Herzog^[12] clearly noted that, in the absence of archaeological evidence or historical documentation, ancient road reconstruction based solely on GIS can hardly transcend hypothetical inference. In recent years this issue has received increasing attention. Common practices include comparing simulated roads with independent control points, site buffer zones, historical corridors, or related remains; some studies also use random points or null models to assess the extent to which simulation results deviate from random distributions^[27–29]. Existing studies based on the Haidai region have applied settlement interaction models and flow-accumulation methods to prehistoric road reconstruction and have attempted validation through training-set and test-set partitioning^[21]. However, such validation typically treats sample partitioning as a technical step and does not systematically distinguish between the different statistical functions of the training set as an input of prior information and the test set as an independent source of evidence. As a result, it remains difficult to determine to what extent model evaluation is constrained by sample size and spatial distribution pattern, rather than by the spatial explanatory power of the road network itself. Accordingly, under conditions of incomplete archaeological data, disentangling the coupled effects of the amount of training information, the amount of testing evidence, and spatial distribution patterns, and further discussing whether the validated road structure possesses interpretability, constitutes the methodological point of departure of this study.

An under-explored issue within existing validation frameworks is how the incompleteness of archaeological records simultaneously affects road generation and road validation. Known sites are not a complete representation of the historical population; their discovery and preservation are affected by multiple factors such as burial, erosion, coastline migration, lacustrine evolution, and survey intensity. Existing samples therefore approximate a non-randomly missing subset rather than a uniform sample of the historical population^[30,31]. This problem is particularly pronounced in coastal and plain regions, where environmental change not only affects settlement distribution itself, but also alters the probabilities of site preservation and discovery^[32–34]. Under such conditions, if limited samples are used simultaneously for road generation and result evaluation, and if the roles of training-set information and test-set evidence are not distinguished, then changes in evaluation metrics become difficult to interpret accurately. Low validation scores may indicate that the path-generation mechanism has failed to capture real connections effectively, but they may also merely reflect a decline in statistical power caused by insufficient sample size or biased sample distribution^[35]. Similarly, if one discusses the road grade–length relationship, network hierarchical structure, or developmental stage without first clarifying sample-size effects, statistical appearances caused by sample sparsity or modeling bias may be misread as real signals of network evolution. For research that seeks to relate the development of road systems, the formation of hierarchies, and the complexification of regional societies, this issue is of fundamental significance.

This issue is especially pertinent to the Haidai region. Located in the eastern part of the lower Yellow River, the Haidai region is a composite zone where mountains, piedmont plains, and coastal lowlands intersect, possessing both strong geographical segmentation and clear corridor

properties and potential for regional interaction. More importantly, the region contains a relatively continuous cultural sequence from Houli, Beixin, and Dawenkou to Longshan and Yueshi, making it an important area for observing prehistoric settlement expansion, regional integration, and the civilizational process in eastern China. Existing research has shown that the Haidai region was not an isolated and self-contained unit of development, but long occupied a junction position among the lower Yellow River, the eastern coast, and the Jianghuai region. According to Luan Fengshi's discussion of the cultural affinities of the Houli culture^[36], several core elements of the early Majiabang culture may have had relatively deep origins in the Houli culture and Xiaoshankou-type remains. Comparisons between the Houli–Beixin system and the Gongsan culture of the Korean Peninsula, as well as some early cylindrical round-bottomed vessel systems in Kyushu, Japan, likewise suggest that the Haidai region may have been one of the key nodes in the prehistoric exchange network along the eastern coast of East Asia. For research on road networks, this means that the Haidai region is suitable not only for discussing internal regional transportation organization, but also for understanding the mechanisms of interregional cultural interaction from the perspective of spatial connectivity.

To address the above issues systematically, this study designs a controlled experimental framework that simultaneously disentangles sample-size effects and spatial-distribution effects. First, taking the relatively high-density site data of the Longshan period in the Haidai region as the baseline, we separately examine the independent driving effects of training-set size and test-set size on the accuracy and robustness of road network evaluation. Second, we introduce the Dawenkou period in the Haidai region and the Liangzhu period in the Taihu Basin as comparative cases to test the influence of different settlement spatial configurations on the applicable

boundaries of the validation framework. Finally, on the basis of validation, we extract the hierarchical structure of the backbone road network and examine whether its length distribution exhibits clearer tendencies toward hierarchy and scaling. The present study thus incorporates road network reconstruction, network validation, and analysis of network developmental characteristics into a single analytical framework.

Results

To evaluate the accuracy and reliability of the reconstructed road networks, this study adopts a training-set and test-set split design. All known site points were divided into two non-overlapping subsets by sparse random sampling: the training set was used exclusively to generate the traffic-flow road network, while the test set did not participate in network generation at any stage. The test set served solely as independent external evidence for evaluating network quality (see Methods for the detailed partitioning procedure and calculation formulas). The following sections examine, in turn, the independent driving effects of test-set size, training-set size, and settlement spatial distribution patterns on the evaluation metrics, and then extract the hierarchical structure of the road network and its associated properties within the validation framework.

Effects of test-set density on the accuracy and robustness of the road network

Among the five periods of Neolithic culture in the Haidai region, the Longshan period represents the peak of civilizational development and has the richest record of known settlements, with 428 sites. This section takes the fixed convex hull formed by the boundaries of Longshan-period sites as the study area and examines the specific driving effects of training-set and test-set density on the reconstruction results.

As shown in Figure 1, under the condition that the proportion of the training set is fixed at 15% ($N_{\text{train}} = 60$), the number of independent control points used for validation was increased stepwise to observe the patterns of change in the Capture Rate (CR) and the Z value with the number of test-set sites. Under a fixed training set, as the size of the test set increases, the Z value rises continuously and exhibits an approximately power-law growth relationship ($Z \propto m^b$, $b \approx 0.5$). In contrast, CR fluctuates markedly when the test set is small and then converges rapidly. This indicates that increasing the size of the test set primarily improves the adequacy of the validation evidence, rather than substantively improving the spatial fitting quality of the road network itself. To verify the generality of these patterns, scatter plots and fitted functions for Capture Rate against number of test-set sites and Z value against Capture Rate were calculated and plotted for training-set proportions ranging from 15% to 90%, with a step size of 5%. To verify the generality of the fitted Z value against number of test-set sites relationship, we calculated the exponent and the coefficient of determination in the fitted function under each training-set proportion (Figure 3a, b). To describe the state shown in Figure 1, in which the fluctuation of the Capture Rate decreases rapidly as the number of test-set sites increases, we divided the horizontal axis of Figure 1 into four approximately equal segments (Figure 2), and then calculated the standard deviation of the vertical coordinates of the scatter points in each segment (the Capture Rate). Four such standard deviations were obtained for each training-set proportion. We then calculated the mean values of these four standard deviations across all training-set proportion scenarios and plotted them in Figure 3c.

As shown in Figure 3a, for different training-set proportions, the exponent b of the fitted function varies within the interval $[0.477, 0.6161]$ at the 95% confidence level, and the theoretical value of

0.5 falls within this interval. Figure 3b shows that the coefficient of determination R^2 varies within the range [0.513, 0.735] at the 95% confidence level, indicating that the significance of the fitted function is relatively stable. Figure 3c shows that, as the number of test-set sites increases, the fluctuation of the Capture Rate decreases rapidly to below $\pm 1\%$.

Effects of training-set size: coupled analysis and isolation of the net effect

When discussing the effect of training-set size, the classical design is to hold the test set constant to observe the impact of changes in the training set. However, under the condition of incomplete archaeological data, holding the test set constant is not the only reasonable option. The analysis above has already shown that once the number of test-set sites reaches a certain level, the estimation of the Capture Rate is no longer sensitive to differences in test-set size, and an increase in the number of test-set sites enables the road network to achieve a better Z-value performance. Under complementary partitioning, as the number of training-set sites increases, the Capture Rate curve exhibits a growth trend toward a saturation upper limit. At the right end of the scatter plot, however, the degree to which the scatter points deviate from the fitted curve increases (Figure 4). This is because an increase in the number of training-set sites leads to a decrease in the number of test-set sites, and a smaller number of test-set sites produces greater fluctuation in the Capture Rate, as shown in Figure 4. In Figure 5, the Z value shows a decreasing trend as the training-set size increases. This does not necessarily indicate a decline in road network quality: while the training-set size increases, the test-set size decreases, and a decrease in test-set size reduces the Z value (as already evident in Figure 1).

To isolate the confounding effect, a further control experiment was conducted under a fixed control-point size of $m = 43$, with an equal number of control points drawn from the remaining

site points for validation under different training-set sizes. The results show that, after the control-point size is fixed, the Capture Rate still increases with training-set size and tends toward saturation (Figure 6), indicating that the primary driving factor of the Capture Rate is the amount of information in the training set rather than variation in control-point size. In contrast, the Z value exhibits some fluctuation when the training set is small, but once the training set reaches a certain size, it rapidly stabilizes and no longer exhibits a systematic monotonic change (Figure 7). This indicates that, after the amount of training information reaches a level of sufficient training, the statistical characteristics of the road-network structure have already become saturated, and further increases in the number of training points make only a limited marginal contribution to Z.

Driving mechanism of the Z value under the complementary condition

As discussed above, once the sizes of the test set and the training set have both reached a certain level, the Capture Rate is determined primarily by the training-set size, whereas the Z value is determined primarily by the test-set size. The decreasing relationship between the Z value and the number of training-set sites shown in Figure 5 is therefore caused by the reduction in the number of test-set sites. To verify this, we replaced the horizontal axis in Figure 5, changing it from the number of training-set sites to the complementary number of test-set sites, thereby obtaining Figure 8.

In Figure 8, although the number of training-set sites decreases as the number of test-set sites increases, the influence of training-set size on the Z value becomes negligible once the training set reaches a certain scale (Figure 7). The experiment shown in Figure 8 therefore effectively constitutes an equivalent controlled-variable experiment: with the number of training-set sites controlled at a level producing approximately the same effect, it examines the influence of the

number of test-set sites on the Z value. In Figure 8, the fitted power-function trend of Z value against number of test-set sites reappears, and its exponent (0.5014) is very close to the value of 0.4920 in Figure 1; both are near the theoretical value of $b \approx 0.5$.

Figure 1 depicts the progressive slices through which the Z value accumulates as the test set increases from small to large under a single, fixed training set. By contrast, each scatter point in Figure 8 represents the limiting situation in which, once a specific training set has been defined, all of the remaining available independent sites are used for validation. The rightmost scatter point in Figure 8 is thus equivalent to the situation at the far right of Figure 1, where the test sample is most sufficient; Figure 8 as a whole represents the collection, across different training-set proportions, of the most sufficient test-set state of a single curve such as that shown in Figure 1.

Spatiotemporal effects of spatial distribution patterns on validation metrics

The validation-decoupling framework established above, in which the Capture Rate is driven by the amount of training information while the Z value is constrained by the amount of test evidence, is based primarily on Longshan-period data with a uniform macroscopic distribution. To examine the spatiotemporal generality of this framework and the boundary conditions of its applicability, this study further introduces the Dawenkou period in the Haidai region and the Liangzhu period in the Taihu Lake basin as comparative validations (Figures 13 and 11). To eliminate differences in total area and base quantity, the density scale (sites per 10,000 km²) is uniformly adopted for horizontal comparison.

To investigate the mechanism underlying the divergence of the fitted curves across periods, we extracted three indicators of spatial pattern from the sites of each period: local clustering (Nearest

Neighbor Index, NNI), macroscopic directionality (flattening ratio of the standard deviational ellipse), and global unevenness (Gini coefficient) (Table 1).

In the relationship between the Capture Rate and training-set density (Figure 9), all three periods exhibit an asymptotic trend of saturating growth, demonstrating that the pattern whereby the Capture Rate is driven by information quantity possesses mathematical-form universality across different times and spaces. To identify an appropriate functional form for describing this saturation, we fitted five candidate models to the complementary-partitioning Capture Rate data ($n = 17$ training-set size levels for the Longshan period): Michaelis-Menten, exponential saturation, Weibull CDF with fixed asymptote ($A = 100\%$), Weibull CDF with free asymptote, and logistic. We performed formal model comparison using the small-sample corrected Akaike Information Criterion (AIC_c). The results are summarized in Table 3.

The three two-parameter models, Michaelis-Menten ($AIC_c = 43.14$, $R^2 = 0.957$), exponential saturation ($\Delta AIC_c = 0.50$), and Weibull ($A = 100\%$, $\Delta AIC_c = 0.59$), are statistically indistinguishable (all $\Delta AIC_c < 2$), and their fitted curves are nearly identical within the data range. The three-parameter models (Weibull with free asymptote, logistic) were penalized by AIC_c for the additional parameter without commensurate improvement in fit, indicating that two parameters are sufficient to describe the saturation. The Michaelis-Menten model was selected for display in Figure 9 because it achieved the lowest AIC_c and its parameters carry direct interpretive value:

$V_{\max} = 113.6 \pm 6.5$ (asymptotic capture rate, %) and $K_m = 177.9 \pm 23.3$ (the training-set size at which half the asymptotic capture rate is reached). The fitted Weibull ($A = 100\%$) parameters are $\lambda = 225.5 \pm 6.8$ (characteristic scale) and $k = 0.791 \pm 0.050$ (shape parameter; $k < 1$ indicates a decelerating approach to saturation without an inflection point). The near-equivalence of all two-

parameter saturation functions confirms that the saturation trend is robust and does not depend on the specific choice of functional form.

However, differences among the periods in the rate of increase of the curves and in their limiting values are correlated with their spatial distribution indicators. For the Dawenkou and Longshan periods, both belonging to the Haidai region, the Gini coefficients differ notably (0.200 vs 0.285).

When the training-set densities fall within the overlapping interval (5–15 sites per 10,000 km²), the basic road networks generated by the Settlement Interaction Model show highly similar capture performance for the two periods. By contrast, the Capture Rate curve for the Liangzhu period rises markedly, exceeding 70% even at low training densities. Liangzhu sites have the lowest Nearest Neighbor Index (0.512) and the highest Gini coefficient (0.594), indicating a strongly polarized core–periphery distribution pattern, with the majority of sites concentrated in micro-topographic core areas. Under this pattern, the model requires only a very small number of training points to generate a high-intensity local network within the core area, thereby readily covering the great majority of test points that are likewise located within that core area.

In the relationship between the Z value and test-set density (Figure 10), the evolution of statistical confidence is sensitive to the assumption of spatial stationarity. Among the three periods, the Liangzhu period has the highest flattening ratio of the standard deviational ellipse (0.686), followed by the Longshan period (0.599), which exhibits a relatively balanced and centripetal macroscopic distribution. This balanced distribution satisfies the first-order stationarity assumption in spatial statistics to a considerable extent, allowing the flow deviation of real points relative to the random background to remain stable, and thereby producing a stable fit to the theoretically expected square-root curve.

Spatial patterns of road networks across the five periods

The extraction-threshold scheme was defined on the basis of Longshan-period data, which contain the largest number of sites, and this flow-classification standard (five hierarchical classes) was then applied to the other four periods, thereby reconstructing the sequence of rise and decline in road-network morphology. In the earliest Houli and Beixin periods, the generated networks consist of sparse, mutually isolated regional closed loops and short branch segments, with no high-order roads appearing. By the Dawenkou period, these developed into two major principal road-network blocks: Hai (the coastal zone) and Dai (the hinterland surrounding Mount Tai), which did not connect with one another. With the transition into the Longshan period, the dense and strongly connected settlement system catalyzed the integration of the Haidai region. By the Yueshi cultural period, however, accompanied by the wave of settlement decline brought about by an abrupt climatic shift, the unified road network disintegrated and contracted into several disconnected subregional networks. (Fig. 13).

Evolution and statistical characteristics of the road length–hierarchy relationship

Figure 11 presents scatter plots of road length against road hierarchy for each cultural period under the five-level classification scheme. Visual inspection suggests that the early periods (Houli and Beixin) may exhibit a broadly linear trend, while the later periods (Dawenkou, Longshan, and Yueshi) show a markedly exponential pattern. However, reliance on R^2 comparison alone is insufficient to establish the superiority of one functional form over another, because a two-parameter exponential model possesses greater curvature flexibility than a two-parameter linear model and will virtually always achieve an equal or higher R^2 for monotonically increasing data.

To resolve this ambiguity, we conducted formal model comparison using the small-sample corrected Akaike Information Criterion (AIC_c) and the Bayesian Information Criterion (BIC). Three candidate models were evaluated for each period and each classification scheme ($k = 3, 4, 5, 6, 7, 8$): linear ($L_r = a + b \cdot r$), exponential ($L_r = a \cdot e^{(b \cdot r)}$), and a three-parameter offset power-law ($L_r = a \cdot r^b + c$). Nonlinear models were fitted via the Levenberg–Marquardt algorithm; the linear model via ordinary least squares. Akaike weights (w) were computed to quantify the relative support for each model. The $k = 5$ results, corresponding to the five-class scheme used throughout this study, are summarized in Table 2.

The AIC_c comparison reveals three key findings. First, linear versus exponential model comparison indicates a temporally staged pattern rather than a single universal functional form. In the Houli period, the linear model is the AIC_c -best model across all six classification schemes ($k = 3-8$), though at $k = 5$ the two models are statistically indistinguishable ($|\Delta AIC_c| = 0.98 < 2$). In the Beixin period, the linear model is preferred in five of six schemes ($k = 3$ favors exponential marginally), with moderate support at $k = 5$ ($\Delta AIC_c = 3.20$, $w_{lin} = 0.832$). However, from the Dawenkou period onward, the exponential model is the best model in 17 of 18 period-by- k combinations (only Dawenkou $k = 3$ favors linear), and it receives decisive support in all cases with $k \geq 4$ ($\Delta AIC_c > 10$). The data therefore support a staged transition: early periods (Houli, Beixin) characterized by linear scaling, and later periods (Dawenkou, Longshan, Yueshi) dominated by exponential hierarchical scaling.

Second, and more importantly, the model selection results follow a clear developmental gradient (Table 2). In the Houli period, the linear model is slightly preferred at $k = 5$ but the two models are statistically indistinguishable ($\Delta AIC_c = -0.98$, $|\Delta AIC_c| < 2$; $w_{lin} = 0.620$, $w_{exp} = 0.380$), and the

best model shifts inconsistently across all six k values (linear preferred in all six, but five have $|\Delta AIC_c| < 2$). The exponential decay parameter b is poorly constrained ($b = 0.311 \pm 0.275$ at $k = 5$; coefficient of variation = 88.5%), meaning the exponential curve can take almost any shape while still fitting the data. In the Beixin period, the linear model is consistently and clearly favored ($\Delta AIC_c = -3.20$ at $k = 5$, $w_{lin} = 0.832$; linear preferred in five of six k values), with the parameter b becoming moderately identifiable ($b = 0.588 \pm 0.166$ at $k = 5$; CV = 28.2%). From the Dawenkou period onward, the exponential model achieves decisive support across all classification levels (all $\Delta AIC_c \geq 23.1$, $w_{exp} \geq 0.999$ for all k), with the parameter b tightly constrained (e.g., $b = 1.821 \pm 0.066$ at Dawenkou $k = 5$; CV = 3.6%).

Third, the three-parameter offset power-law model was never selected as the best model by AIC_c . Its offset parameter c was estimated as approximately zero ($c \approx 10^{-15}$ to 10^{-5}) with infinite standard error in nearly all fits, and the exponent b frequently hit the optimization upper bound of 10, indicating that the model is overparameterized for data with only $k = 3-8$ points per period. The data therefore provide no evidence that adding an intercept term improves upon the two-parameter exponential description.

To further assess robustness, we examined whether the coefficient of determination (R^2) of the exponential fit remains stable when the number of hierarchy classes (k) is varied from 3 to 8 (Figure 12). For the underdeveloped early network stages (Houli and Beixin periods), excessively increasing the number of classification levels actually reduces the significance of the fit, with R^2 values exhibiting a declining and oscillating pattern. However, for the mature Dawenkou, Longshan, and Yueshi periods, regardless of how the number of class gradations is adjusted, the fitted results consistently display excellent statistical consistency and an exponential character.

We also note an inherent limitation: model selection is performed on $k = 3-8$ data points per period (one mean road length per hierarchy class), which constrains statistical power. The AICc formula includes a small-sample correction term $(2K(K+1)/(n-K-1))$ that partially addresses this, but the ability to discriminate between models remains limited when k is small. This limitation most affects the Houli period, where sparse data and weak hierarchical differentiation compound one another.

The results reveal a systematic decoupling between the two evaluation dimensions, accuracy and statistical reliability, under controlled experimental conditions. The near-square-root power-law growth of the Z value with increasing test-set size and the convergence of Capture Rate fluctuation are not artifacts of a specific parameter setting, but general patterns that occur across different training-set proportions. This feature arises from the interaction between the physical process of spatial flow accumulation and classical statistical testing theory.

In the Settlement Interaction Model used in this study, once the basic road network has been generated from a given training set, the spatial distribution and topological structure of high-flow traffic corridors are fixed. Because the actual distribution of prehistoric sites is not spatially random, but is instead strongly constrained by the friction cost of natural geography and tends to cluster along traffic corridors, as the number of independent test points (m) increases, the absolute difference between the total flow captured by the actual test group and the expected flow of random points under the null model, the numerator term in the Z -value calculation, increases approximately linearly with m ^[35,45]. At the same time, according to the Central Limit Theorem, the standard deviation of the flow distribution under the null model (the denominator term) increases with the square root of the sample size^[46]. The ratio of a linear numerator to a square-

root denominator mathematically constrains the overall Z value to follow the square-root law, with a power-function exponent of $b \approx 0.5$.

At the spatial-archaeological level, this phenomenon confirms Cohen's theory of statistical test power^[35]: when the true degree of spatial deviation specified by the model is fixed, the statistical confidence for rejecting the null hypothesis (the assumption of a random distribution) accumulates with the square root of the test sample size. This mechanism, by which statistical power strictly follows a power function of sample size, constitutes the fundamental mathematical reason why the Z value can be used as a quantitative proxy for test sufficiency.

These findings establish what we term a conservative evaluation principle. Because both the Capture Rate and the Z value are monotonic functions of sample size within their respective regimes (CR approaches a saturation asymptote with increasing training-set size, Figure 8; Z grows with the square root of test-set size, Figure 3), validation conducted on currently known sites represents the most pessimistic assessment the model will ever face. As more sites are discovered and incorporated into the analysis, both evaluation metrics can only improve or remain stable. For the practical question of whether a reconstructed prehistoric road network is reliable enough, this principle provides a clear statistical rationale: if a network passes validation with the limited data currently available, its true quality, measured against the unknown complete settlement system, is unlikely to be worse, and is more likely better. The conservative evaluation principle thus transforms a methodological vulnerability (incomplete archaeological data) into an epistemic strength: it provides a lower-bound guarantee that supports the credibility of reconstruction methods even when the full settlement record is unavailable.

In addition, the pattern in which the Capture Rate first fluctuates sharply and then rapidly converges as the number of test-set sites increases (Figures 2 and 3c) reflects the convergence of spatially heterogeneous sampling to the Law of Large Numbers. Prehistoric settlements, especially in the Longshan period, exhibit strong spatial autocorrelation and clustered distribution at the macrogeographical scale. When the test-set size is small, spatial sampling is prone to producing cluster bias^[30]: the accidental selection of settlement clusters located along trunk roads or of remote settlements leads to large fluctuations in the estimation of the Capture Rate (the high standard deviation in the Q1 stage of Figure 3c).

As the validation sample size continues to increase, if the inclusion of site points within the road buffer is regarded as a series of Bernoulli trials, the sample hit ratio will probabilistically converge to the true expected value determined by the training network in that run, and its standard deviation will shrink proportionally^[47]. This indicates that merely increasing the size of the test set cannot substantially improve the model's fitting accuracy to the overall population, but only smooths spatial sampling noise, thereby causing the evaluation indicators to converge toward the true explanatory level of that road network.

The controlled experiments isolating training-set and test-set effects demonstrate a functional decoupling: once the training set has reached a sufficient size, the principal source of variation in the Z value is the test-set size (the amount of validation evidence), rather than the training-set size itself. Conversely, the Capture Rate is dominated by the amount of training information and approaches a saturation limit. This decoupling carries important practical implications for archaeological work. It means that, under conditions of incomplete data, validation assessment must distinguish between two independent dimensions: the physical goodness of fit (represented

by the Capture Rate), which is constrained by the structural conditions of the model-generation stage, and the statistical confidence (represented by the Z value), which is constrained by test power at the validation stage.

Critically, the monotonic nature of both relationships, CR saturating with training-set size and Z accumulating with test-set size, underpins the conservative evaluation principle identified above.

Because both metrics move only in the favorable direction as sample sizes increase, evaluations based on currently available sites serve as a statistical lower bound: if a reconstructed network demonstrates acceptable CR and Z values with the incomplete data at hand, its performance against the true, unobservable settlement system can be expected to be at least as good.

Cross-regional and cross-period comparisons further reveal that validation metric evolution is constrained by the macroscopic spatial distribution of settlements. The Capture Rate growth pattern is dominated by the amount of training information, but its rate of increase and threshold are regulated by local clustering intensity (NNI) and the degree of spatial polarization as indicated by the Gini coefficient. For the Dawenkou and Longshan periods, which share relatively low Gini coefficients (0.200 and 0.285, relative to Liangzhu's 0.594), the basic road networks show closely similar capture performance at overlapping training densities. By contrast, the Liangzhu period's core-periphery distribution allows even sparse training points to generate high-intensity local networks that cover the majority of clustered test points, producing an elevated Capture Rate curve overall. This indicates that the rate of increase in the Capture Rate is influenced not only by geographic friction costs, but also by the modulation of local settlement clustering and global unevenness.

Formal model comparison (Table 3) further demonstrates that the saturation pattern is robust across functional forms: the three two-parameter models (Michaelis-Menten, exponential saturation, Weibull with fixed asymptote) are statistically indistinguishable ($\Delta AIC_c < 0.6$). This confirms that the qualitative conclusion, namely, that Capture Rate approaches a saturation limit with increasing training-set size, does not depend on the specific choice of fitting function.

Whether the Z value stably follows the square-root accumulation law, and therefore whether the conservative evaluation principle holds in its simplest form, depends on the macroscopic form of site distribution. The Longshan period's relatively balanced and centripetal macroscopic distribution satisfies the first-order stationarity assumption in spatial statistics, allowing a stable fit to the theoretical square-root curve and thereby fulfilling the conditions for the conservative principle.

For the Dawenkou period's pronounced belt-like extension (low flattening ratio, 0.438, of the standard deviational ellipse), spatial heterogeneity and effect-size contraction caused by insufficient training points can interfere with the stability of the test statistic. This defines a first boundary condition: when site distribution is strongly directional and training data are sparse, the monotonic accumulation of statistical confidence may be disrupted by network fragmentation. For the Liangzhu period's pronounced core concentration (Gini coefficient = 0.594), sampling variance overwhelms the gradual accumulation of confidence. This defines a second boundary condition: under extreme spatial polarization, the conservative principle is not violated (the network may still be of adequate quality), but the Z value loses its predictable relationship with sample size, making the lower-bound guarantee quantitatively uninformative.

These findings indicate that, before applying the conservative evaluation principle in a given study area, quantitatively assessing the spatial distribution pattern of sites (using the NNI and the standard deviational ellipse) is a necessary preliminary step for determining whether the principle's conditions of applicability are met. When evaluating road networks under incomplete data conditions, researchers should not pursue the absolute magnitude of the Z value in isolation from sample size. Given that the Z value is strictly constrained by the upper limit of the size of the independent test set ($Z \propto m^{0.5}$), one should first assess whether the currently available validation sample size possesses the basic statistical power required to cross the significance threshold. This provides a foundational mathematical basis for objectively understanding test performance in data-poor environments.

Turning to the hierarchical structure of the validated road networks, the formal AICc-based model comparison (Table 2) reveals a developmental gradient from linear to exponential hierarchical length scaling, rather than a binary switch between functional forms.

In the Houli period, characterized by the smallest number of sites and sparsest density, the linear model is slightly preferred but statistically indistinguishable from the exponential model ($k = 5$: $\Delta AIC_c = -0.98$, $|\Delta AIC_c| < 2$), and the exponential decay parameter b is effectively unidentifiable ($CV = 88.5\%$ at $k = 5$). This indicates that, at the earliest stage of network development, the hierarchical differentiation of roads is not yet a statistically robust signal; it is present but near the noise floor, and the hierarchy classes produced by the Jenks algorithm may primarily reflect the physical hierarchy of topographic channels (trunk valleys versus tributary valleys) rather than functionally distinct road classes. The Beixin period shows a clearly transitional state: the linear model gains moderate support at $k = 5$ ($\Delta AIC_c = -3.20$, $w_{lin} = 0.832$), yet the exponential model

already achieves $R^2 = 0.902$, and the parameter b becomes moderately identifiable ($CV = 28.2\%$ at $k = 5$), indicating that road-network hierarchical differentiation has begun to emerge as a statistically recognizable signal.

From the Dawenkou period onward, once settlement density exceeded a critical threshold, exponential hierarchical scaling is unequivocally established: all $\Delta AIC_c > 23$ and Akaike weights > 0.999 for all $k = 3-8$, with the parameter b tightly constrained ($CV < 10\%$), indicating that the hierarchical structure is a robust property of the data rather than a by-product of the classification scheme. This developmental gradient, from linear to exponential scaling, is consistent with the theoretical expectation that hierarchical self-organization in spatial networks emerges gradually with increasing node density and interaction intensity, while also highlighting that the early sparse networks were still dominated by the primordial constraints of natural geographic channel systems and lacked mature functional class differentiation.

The exponential functional form itself is theoretically expected for hierarchically organized branching networks. According to central place theory in geography^[49] and Horton's law in fluvial geomorphology^[50], in a mature transport network each descending hierarchy level serves a geometrically expanding catchment area, such that the total road length at level r follows $L_r \propto e^{(b \cdot r)}$, where $b = \ln(RL)$ and RL is the characteristic length ratio between adjacent levels. The widespread presence of this exponential distribution and its decisive statistical support from the Dawenkou period onward demonstrate that, after settlement density crossed a critical threshold, the Haidai region had formed an ordered complex road system possessing structural self-similarity^[25]. The three-parameter offset power-law model ($L_r = a \cdot r^b + c$) was tested as a potential unified alternative that could, in principle, fit both linear-like ($b \approx 1$, $c > 0$) and exponential-like

regimes, but it was never selected by AIC_c and its offset parameter was consistently unidentifiable, indicating that the data do not require an additional degree of freedom beyond the two-parameter exponential model.

This study clarifies the mechanism-level decoupling of the road-network validation system and identifies a conservative evaluation principle governing network assessment under incomplete archaeological data. The Capture Rate represents the model's physical goodness of fit, is dominated by the amount of training information, and exhibits a saturation limit. The Z value represents the strength of deviation from randomness and, once road-network training is sufficient, is driven primarily by the amount of test evidence, strictly following the square-root accumulation law in statistical inference.

Because both metrics are monotonic functions of sample size, validation conducted on currently known sites establishes a statistical lower bound on true network quality: if a reconstructed network passes validation with limited data, its actual performance is unlikely to be worse. Cross-regional and cross-period comparisons further delineate the boundary conditions of this principle: local clustering, macroscopic directionality, and global unevenness constrain the conditions under which the conservative guarantee holds in its simplest form.

Furthermore, after removing sampling interference through the quantitative validation framework, the road network of the Haidai region is found to exhibit a developmental gradient in the statistical robustness of exponential hierarchical scaling, from weakly expressed and fragile in the earliest period (Houli), through clearly established (Beixin), to unequivocally robust in mature networks (Dawenkou through Yueshi). This suggests that hierarchical self-organization in prehistoric

transport systems is not an all-or-nothing property, but a structural signal that strengthens as networks grow in density and complexity.

Several limitations should be acknowledged. First, site data are affected by modern burial environments (such as alluviation in plains and denudation in hilly areas), and non-random site loss can introduce spatial autocorrelation bias.

Second, the current generation of flow-based road networks relies on static topographic models, whereas the frequent avulsion of river systems and changes in coastlines during the Neolithic period are difficult to correct within such models, thereby constituting a systematic physical error in topological extraction.

Third, the formal model comparison for the road length–hierarchy relationship is performed on $k = 3\text{--}8$ data points per period, which limits statistical power; results for the sparsest periods, especially Houli, should be interpreted with appropriate caution.

Fourth, the saturation-function model comparison for the Capture Rate (Table 3) is based on $n = 17$ training-set size levels from complementary partitioning of Longshan-period data. While this sample size is adequate for distinguishing two-parameter from three-parameter models, the statistical power to discriminate among alternative two-parameter forms is limited. The conclusion that the top three models are indistinguishable ($\Delta AIC_c < 0.6$) should be interpreted in this light.

Future research could incorporate agent-based modeling to provide supplementary logical support from the dimension of human social interaction, and could apply alternative hierarchy-extraction methods (such as quantile classification or head/tail breaks) to further test the robustness of the exponential scaling pattern.

The generation of prehistoric roads is based on friction surfaces and least-cost mechanisms. A composite spatial friction surface was constructed from elevation (DEM), slope, and the distribution of modern and ancient hydrology within the study area. With known archaeological site points as nodes, the SIM radiation model was used to simulate the population and resource interaction gravity between nodes, and the flow-accumulation algorithm was employed to generate an anisotropic traffic-flow raster on the friction surface. By extracting linearly connected traces where flow was concentrated, the basic simulation result of the prehistoric macroscopic transportation network was obtained. For the detailed procedures of physical-friction assignment and traffic-flow road-network generation, see Yong et al.^[21].

To generate and evaluate the traffic-flow road network, the entire set of site points was divided into a training set and a test set. The training set was used to generate the traffic-flow road network, whereas the test set did not participate at all in the generation of the road network and was used only to evaluate the quality of the road network. To avoid the influence of the spatial clustering of site points, both the training set and the test set were selected proportionally from the entire set of site points by sparse random sampling.

Study areas and materials

This paper selects the Taihu Basin as a reference study area for cross-regional comparison. The Taihu Basin is located in the southern plain region of the Yangtze River Delta in China, bordering the East China Sea to the east, adjoining Hangzhou Bay and the area north of the Qiantang River to the south, extending westward to the Maoshan range, the Ningzhen range, and the Tianmu mountains, and reaching northward to the southern bank of the Yangtze River. Centered on Lake

Taihu, it exhibits a dish-shaped geomorphological structure characterized by low and flat terrain in the middle with slightly higher elevations around the periphery (Fig. 14).

Compared with the Haidai region, the Taihu Basin has less topographic relief but a denser network of rivers and lakes; the distribution of settlements and regional interaction are more readily influenced jointly by microtopography, wetland environments, and hydrological conditions^[37].

The Taihu Basin is an important cultural core area of the Neolithic in the lower Yangtze region, where a cultural developmental sequence of Majiabang–Songze–Liangzhu–Maqiao has been established. This paper selects the Liangzhu period within this sequence as the object of cross-regional comparison, primarily because sites of this period are strongly clustered and display pronounced spatially uneven distribution, providing a contrast with the more balanced or belt-like samples of the Haidai region and thereby testing the influence of different spatial patterns on the validation results of road networks^[37].

Data for archaeological sites were primarily derived from the Atlas of Chinese Cultural Relics, including the volumes for Shandong, Jiangsu, Anhui, Hebei, and Henan, as well as from published excavation reports^[38–42]. The detailed information for each site was recorded, including site number, date, area, longitude, and latitude. The digital elevation model (DEM) was obtained from a joint publication of the Ministry of Economy, Trade and Industry of Japan (METI) and the National Aeronautics and Space Administration (NASA), specifically the 30 m SRTM (Shuttle Radar Topography Mission) data (<http://www.gscloud.cn>). In addition, the original data on land cover and water systems were derived from the 1985 data on various land-cover types in the CLCD dataset of Wuhan University^[43].

The overall research design and analytical framework of this study are illustrated in Fig. 15. The following subsections describe each component in detail.

Flow model for prehistoric road network reconstruction and evaluation of road network quality

The generation of prehistoric roads is based on friction surfaces and least-cost mechanisms. A composite spatial friction surface was constructed from elevation (DEM), slope, and the distribution of modern and ancient hydrology within the study area. With known archaeological site points as nodes, the SIM radiation model was used to simulate the population and resource interaction gravity between nodes, and the flow-accumulation algorithm was employed to generate an anisotropic traffic-flow raster on the friction surface. By extracting linearly connected traces where flow was concentrated, the basic simulation result of the prehistoric macroscopic transportation network was obtained. For the detailed procedures of physical-friction assignment and traffic-flow road-network generation, see Yong et al.^[21].

Following Frachetti^[15] and Yong et al.^[21], we used the Capture Rate and the Z value of the t-test to evaluate the accuracy and robustness of the road network. Their calculation methods are given in Equations (1) and (2):

$$CR = \frac{N_{hit}}{N_{test}} \quad (1)$$

$$Z = \frac{CR_{obs} - \mu_{null}}{\sigma_{null}} \quad (2)$$

Here, the Capture Rate (CR) is the proportion of test-set site points that fall within the road buffer zone, and the Z value characterizes the strength of the non-random deviation of the real test-point set relative to the random null model. N_{hit} is the number of test-set archaeological site points that fall within the 500 m buffer zone of the road network, and N_{test} is the total number of

archaeological site points in the test set. CR_{obs} is the capture rate of the real test set, while μ_{null} and σ_{null} are the mean and standard deviation, respectively, of the capture-rate distribution under the random null model. The number of random site-point sets is equal to that of the test set, with 200 groups, which were randomly generated within the convex hull formed by the entire set of site points by means of sparse random sampling.

To ensure the independence of the evaluation results, previous studies^[15,21] have adopted the procedure of separating the test-point set from the training-point set so that it does not participate in road network generation. However, they have not discussed in depth what scale the training set and the test set need to reach, nor the effects of their sizes on the quality-evaluation parameters (CR and Z). In actual archaeological contexts, the number of site points discovered is always smaller than the number of sites that once existed. This means that the number of points in the training set participating in road reconstruction and the number of points in the test set participating in road-network quality evaluation are always smaller than the numbers that should have participated under the true situation. However, the effect of this missingness on the quality of the true road network has not been sufficiently discussed.

To examine the effects of the numbers of archaeological site points in the training set and the test set on CR and the Z value, one variable must be held constant to observe the effect of the other on road-network quality evaluation.

Extraction of road-network hierarchy and examination of scaling relationships

After completing the quality evaluation, we extract the hierarchical structure of the validated traffic-flow road network to analyze organizational characteristics and scaling relationships. The

purpose of this analysis is to identify how the network is organized, rather than to determine whether it is effective.

First, for the continuous numerical surface formed by the flow-accumulation results, we use the Jenks Natural Breaks method to classify road hierarchy. This method determines class boundaries according to the natural clustering structure within the data and converts continuous flow values into discrete physical classes while minimizing within-class differences and maximizing between-class differences, thereby reducing the subjectivity introduced by artificially specified thresholds.

Second, after completing the hierarchical classification, to avoid route-length errors caused by diagonal connections during the vectorization of raster roads, we extract refined single-line vectors and calculate the actual cumulative length of roads at each hierarchical level. Let the total length of roads at level r be denoted as L_r ; this then constitutes the basic analytical series of road hierarchy against road length for subsequent examination of scaling relationships.

Third, to identify whether road networks in different cultural periods exhibit stable hierarchical organizational patterns, we compare model fitting for the relationship between road hierarchy r and total road length. Given that early networks and mature networks may correspond to different degrees of hierarchical differentiation, this study examines both the linear model and the exponential model:

$$L_r = a + b \cdot r \quad (3)$$

$$L_r = a \cdot e^{b \cdot r} \quad (4)$$

where a and b denote the parameters to be estimated. To move beyond simple R^2 comparison, we conducted formal model selection using the small-sample corrected Akaike Information Criterion (AIC_c):

$$AIC_c = n \cdot \ln \frac{RSS}{n} + 2K + \frac{2K(K+1)}{n-K-1} \quad (5)$$

where n is the number of hierarchy levels (k), K is the number of model parameters ($K = 2$ for linear and exponential), and RSS is the residual sum of squares. We also computed the Bayesian Information Criterion (BIC) for corroboration. Akaike weights (w_i) were calculated to quantify the conditional probability that each model is the best among the candidate set, given the data. We selected AIC_c as the primary criterion over alternative approaches (such as Vuong's test for non-nested models or likelihood-ratio tests) for two reasons: (i) the candidate models are non-nested and have equal numbers of parameters (linear vs. exponential, $K = 2$), so likelihood-ratio tests are not directly applicable; (ii) the sample size per period ($n = k = 3-8$) is small, conditions under which AIC_c 's small-sample correction term is particularly appropriate, whereas the asymptotic assumptions of Vuong's test may not be met. Following standard conventions^[52], models with $\Delta AIC_c \leq 2$ are considered to have substantial support, $2 < \Delta AIC_c \leq 4$ considerably less support, $4 < \Delta AIC_c \leq 7$ weak support, $7 < \Delta AIC_c \leq 10$ strong support, and $\Delta AIC_c > 10$ essentially no support relative to the best model.

In addition to the two-parameter models, we tested a three-parameter offset power-law model:

$$L_r = a \cdot r^b + c \quad (6)$$

where a , b , and c are estimated parameters ($K = 3$). This model serves as a flexible alternative that could, in principle, approximate both linear-like behavior ($b \approx 1$ with $c > 0$) and exponential-like curvature ($b > 1$). Models (4) and (6) were fitted via nonlinear least squares using the Levenberg–Marquardt algorithm; model (3) via ordinary least squares.

To further examine whether the scaling relationship depends on any arbitrary classification scheme, we vary the number of road hierarchical levels from 3 to 8 and systematically investigate

changes in model selection results. If the best-fitting model and its statistical support for a given period remain consistent under different settings for the number of hierarchical levels, this indicates that its road hierarchical structure is robust; conversely, if the optimal model or its evidence strength fluctuates with k , this suggests that the hierarchical organization has not yet fully developed, or that the classification scheme exerts a non-negligible influence on the apparent structure.

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code availability

The code generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgements

This work was supported by the Natural Science Foundation of Guangdong Province (2024A1515012440), the Education Science Planning Project of Guangdong Province (Higher Education Special Project, 2025GXJK0555), the Teaching Reform Project of Huizhou University (No. 1702242050), and the Guangdong Provincial Higher Education Teaching Reform Project (No. 1702241020).

Author contributions

B.W. conceived the framework of the study. Q.Z. wrote the manuscript. Y.H. and H.G. developed the code and performed the main computational analyses. Z.Z., H.Y., and D.F. refined the research methodology, including the statistical reliability analysis. J.C., R.Z. and Z.K. collected the data and produced the figures.

Competing interests

The authors declare no competing interests.

References

- [1] Bell, T. & Lock, G. Topographic and cultural influences on walking the Ridgeway in later prehistoric times. in *Beyond the Map: Archaeology and Spatial Technologies* (ed. Lock, G.) 85–100 (IOS Press, 2000).
- [2] Knappett, C. *An Archaeology of Interaction* (Oxford University Press, 2011).
- [3] Brughmans, T. Thinking Through Networks: A Review of Formal Network Methods in Archaeology. *J. of Archaeol. Method and Theory* **20**, 623–662 (2013).
- [4] Collar, A., Coward, F., Brughmans, T. & Mills, B. J. Networks in Archaeology: Phenomena, Abstraction, Representation. *J. of Archaeol. Method and Theory* **22**, 1–32 (2015).
- [5] Franklin, K. Archaeology of the Silk Road: Challenges of Scale and Storytelling. *J. of Archaeol. Res.* **32**, 263–308 (2024).
- [6] Evans, T. S., Rivers, R. J. & Knappett, C. Interactions In Space For Archaeological Models. *Advances. in Complex Syst.* **15**, 1150009 (2012).

- [7] Verhagen, P., Nuninger, L. & Groenhuizen, M. R. Modelling of pathways and movement networks in archaeology: an overview of current approaches. in *Finding the Limits of the Limes* 217–249 (Springer, 2019).
- [8] Fábrega-Álvarez, P. & Parcero-Oubiña, C. Proposals for an archaeological analysis of pathways and movement. *Archeol. Calc.* **18**, 121–140 (2007).
- [9] Howey, M. C. Using multi-criteria cost surface analysis to explore past regional landscapes: a case study of ritual activity and social interaction in Michigan, AD 1200–1600. *J. of Archaeol. Sci.* **34**, 1830–1846 (2007).
- [10] Fiz, I. & Orengo, H. Simulating communication routes in Mediterranean alluvial plains. in *Layers of Perception: Proc. 35th Int. Conf. CAA* (eds Posluschny, A., Lambers, K. & Herzog, I.) (Habelt, 2007).
- [11] Gietl, R., Doneus, M. & Fera, M. Cost distance analysis in an alpine environment: comparison of different cost-surface modules. in *Layers of Perception: Proc. 35th Int. Conf. CAA* (eds Posluschny, A., Lambers, K. & Herzog, I.) (Habelt, 2007).
- [12] Herzog, I. The potential and limits of optimal path analysis. in *Computational Approaches to Archaeological Spaces* 179–211 (Left Coast Press, 2014).
- [13] Güimil-Fariña, A. & Parcero-Oubiña, C. “Dotting the joins”: a non-reconstructive use of Least Cost Paths to approach ancient roads. The case of the Roman roads in the NW Iberian Peninsula. *J. of Archaeol. Sci.* **54**, 31–44 (2015).
- [14] White, D. A. & Barber, S. B. Geospatial modeling of pedestrian transportation networks: a case study from precolumbian Oaxaca, Mexico. *J. of Archaeol. Sci.* **39**, 2684–2696 (2012).

- [15] Frachetti, M. D., Smith, C. E., Traub, C. M. & Williams, T. Nomadic ecology shaped the highland geography of Asia's Silk Roads. *Nature* **543**, 193–198 (2017).
- [16] Beyin, A., Hall, J. & Day, C. A. A Least Cost Path Model for hominin dispersal routes out of the East African Rift region (Ethiopia) into the Levant. *J. of Archaeol. Science: Rep.* **23**, 763–772 (2019).
- [17] Tobler, W. *Three Presentations on Geographical Analysis and Modelling* (National Center for Geographic Information and Analysis, 1993).
- [18] Field, S., Glowacki, D. M. & Gettler, L. T. The Importance of Energetics in Archaeological Least Cost Analysis. *J. of Archaeol. Method and Theory* **30**, 363–396 (2023).
- [19] Llobera, M. Memory at your feet: Modeling the agency of past trails. *J. of Archaeol. Science: Rep.* **29**, 102177 (2020).
- [20] Howey, M. C. Multiple pathways across past landscapes: circuit theory as a complementary geospatial method to least cost path for modeling past movement. *J. of Archaeol. Sci.* **38**, 2523–2535 (2011).
- [21] Yong, H. et al. Reconstructing and tracing the evolution of the road networks in the Haidai region of China during the Bronze and Early Iron Ages. *Herit. Sci.* **12**, 145 (2024).
- [22] Barabási, A. & Albert, R. Emergence of Scaling in Random Networks. *Sci.* **286**, 509–512 (1999).
- [23] Barthélemy, M. Spatial networks. *Phys. Rep.* **499**, 1–101 (2011).
- [24] Radivojević, M. & Grujić, J. Community structure of copper supply networks in the prehistoric Balkans: An independent evaluation of the archaeological record from the 7th to the 4th millennium BC. *J. of Complex Netw.* **6**, 106–124 (2018).
- [25] Batty, M. The Size, Scale, and Shape of Cities. *Sci.* **319**, 769–771 (2008).

- [26] Strano, E., Nicosia, V., Latora, V., Porta, S. & Barthélemy, M. Elementary processes governing the evolution of road networks. *Scientific. Rep.* **2**, 296 (2012).
- [27] Khamsiri, S., Venunan, P., Khaokheiw, C., Silapanth, P. & Pailoplee, S. Reconstructing the ancient route network in the Thailand–Cambodia borders: a case study of the Angkorian Royal Road. *Archaeol. Res. Asia* **27**, 100302 (2021).
- [28] Marciak, M., Sobczyński, D. & Čilová, M. In search of ancient pre-Roman imperial roads: a case study of the application of remote sensing in road archaeology in the Southern Levant. *Remote Sens.* **15**, 3237 (2023).
- [29] Groenhuizen, M. R. & Verhagen, P. Testing the Robustness of Local Network Metrics in Research on Archeological Local Transport Networks. *Front. in Digit. Humanit.* **3** (2016).
- [30] Orton, C. *Sampling in Archaeology* (Cambridge University Press, 2000).
- [31] Surovell, T. A. & Brantingham, P. J. A note on the use of temporal frequency distributions in studies of prehistoric demography. *J. of Archaeol. Sci.* **34**, 1868–1877 (2007).
- [32] Wang, Q. Burial depth of Holocene paleocoastlines on the west coast of Bohai Bay. *Quat. Sci.* **24**, 329–335 (2004). [in Chinese].
- [33] Chen, Z. & Wang, Z. The Neolithic in the Taihu Basin: water and the rise and fall of culture. *Science* **50**, 29–32 (1998). [in Chinese].
- [34] Zhu, L. et al. Cultural development of the Liangzhu period and sea-level change. *Prog. Geogr.* **30**, 121–128 (2011). [in Chinese].
- [35] Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Routledge, 2013).
- [36] Luan, F. S. *Archaeological Research in the Haidai Region* (Shandong University Press, 1997). [in Chinese].

- [37] Yang, L., Li, H., Zhang, S., Zhao, Y. & Sheng, Y. Cultural evolution of Neolithic archaeology in the Taihu Lake region of China based on complex network model. *npj. Herit. Sci.* **13**, 117 (2025).
- [38] State Administration of Cultural Heritage. *Atlas of Chinese Cultural Relics: Shandong Fascicule* (Sinomap Press, 2007).
- [39] State Administration of Cultural Heritage. *Atlas of Chinese Cultural Relics: Jiangsu Fascicule* (Sinomap Press, 2008).
- [40] State Administration of Cultural Heritage. *Atlas of Chinese Cultural Relics: Anhui Fascicule* (Sinomap Press, 2014).
- [41] State Administration of Cultural Heritage. *Atlas of Chinese Cultural Relics: Hebei Fascicule* (Cultural Relics Publishing House, 2013).
- [42] State Administration of Cultural Heritage. *Atlas of Chinese Cultural Relics: Henan Fascicule* (Sinomap Press, 1991).
- [43] Yang, J. & Huang, X. The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019. *Earth System. Sci. Data* **13**, 3907–3925 (2021).
- [44] Jenks, G. F. The data model concept in statistical mapping. *Int. Yearb. Cartogr.* **7**, 186–190 (1967).
- [45] Casella, G. & Berger, R. L. *Statistical Inference* 2nd edn. (Duxbury, 2002).
- [46] Rogerson, P. *Statistical Methods for Geography* 5th edn. (SAGE Publications, 2019).
- [47] Agresti, A. *Categorical Data Analysis* 3rd edn. (John Wiley & Sons, 2013).
- [48] Bevan, A. & Wilson, A. Models of settlement hierarchy based on partial evidence. *J. of Archaeol. Sci.* **40**, 2415–2427 (2013).
- [49] Christaller, W. *Die Zentralen Orte in Süddeutschland* (Gustav Fischer, 1933).

- [50] Horton, R. E. Erosional Development Of Streams And Their Drainage Basins; Hydrophysical Approach To Quantitative Morphology. *Geol. Soc. of America. Bull.* **56**, 275 (1945).
- [51] Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
- [52] Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference*. (Springer New York, 2004).

Tables

Table 1. Comparison of the core indicators of spatial distribution patterns of site points across periods.

Cultural Period	Total Sites	NNI	Ellipse Flattening Ratio	Gini Coefficient
Dawenkou	225	0.764	0.438	0.200
Longshan	428	0.660	0.599	0.285
Liangzhu	273	0.512	0.686	0.594

Table 2. Formal model comparison (AICc) of linear vs. exponential fits to the road length–hierarchy relationship at k = 5 hierarchy levels, with the Akaike weight for the linear model and the range of ΔAIC_c (linear minus exponential) observed across k = 3–8.

Period	k	Best Model (AIC _c)	AIC _{c_li} n	AIC _{c_ex} p	ΔAIC_c (lin – exp)	w _{lin}	ΔAIC_c range	Evidence Grade:
Houli	5	linear (indistinguishable)	42.84	43.81	-0.98	0.620	-2.8 – - 0.8	Insignificant ($ \Delta AIC_c < 2$)
Beixin	5	linear	38.44	41.64	-3.20	0.832	-3.2 – - 0.7	Weak
Dawenkou	5	exponential	93.38	60.27	33.11	0.000	28.4 – 39.6	Decisive
Longshan	5	exponential	101.95	67.03	34.92	0.000	32.7 – 44.7	Decisive
Yueshi	5	exponential	84.13	59.04	25.09	0.000	24.1 – 37.0	Decisive

Notes: $\Delta AIC_c (\text{lin} - \text{exp}) = AIC_c_{\text{linear}} - AIC_c_{\text{exponential}}$; negative values favor the linear model, positive

values favor the exponential model. Evidence grades follow standard conventions: ΔAIC_c 0–2 = insubstantial, 2–4

= weak, 4–7 = moderate, 7–10 = strong, >10 = decisive. The offset power-law model ($L_r = a \cdot r^b + c$) was

included in all comparisons but was never selected as the best model by AIC_c .

Table 3. Model comparison of saturation functions fitted to the Capture Rate vs. training-set

size data (Longshan period, complementary partitioning, n = 17).

Model	K	RMSE	R ²	AIC _c	ΔAIC_c	Akaike weight:
Michaelis-Menten	2	3.084	0.9570	43.14	—	0.3531
Exponential saturation	2	3.130	0.9557	43.65	0.51	0.2736
Weibull (A=100)	2	3.138	0.9555	43.73	0.59	0.2629
Weibull (free A)	3	3.087	0.9569	46.17	3.03	0.0776
Logistic	3	3.247	0.9523	47.89	4.75	0.0328

Notes: K = number of estimated parameters. ΔAIC_c is the difference from the best model (Michaelis-Menten).

Akaike weights represent the conditional probability that each model is the best among the candidate set, given the

data. Evidence grades follow Burnham and Anderson (2002)^[52]: $\Delta AIC_c < 2$ = substantial support, 2–4 =

considerably less support, 4–7 = weak support, 7–10 = strong support, >10 = essentially no support. All two-

parameter models are statistically indistinguishable ($\Delta AIC_c < 0.6$).

Figures

Figure 1. Relationship between the Z value and the number of control points (test-set sample size) under a fixed training-set size ($N_{\text{train}} = 60$)

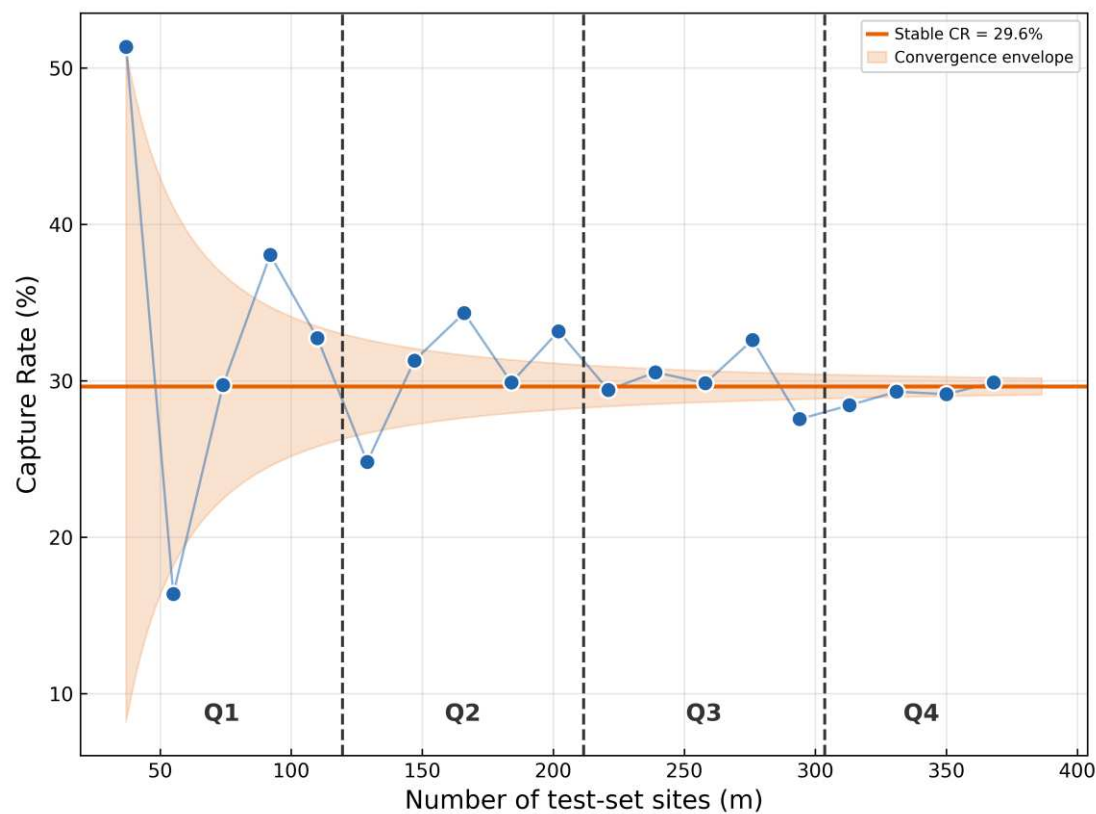


Figure 2. Relationship between the Capture Rate and the number of test-set sites under a fixed training-set size ($N_{\text{train}} = 60$).

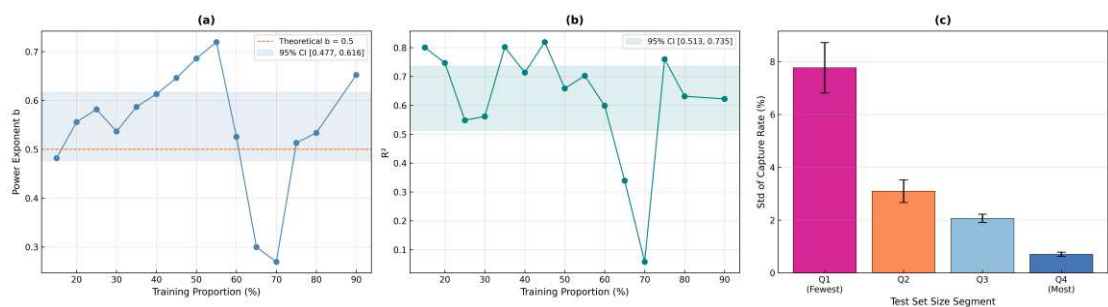


Figure 3. Fitted parameters of the Z value against number of test-set sites relationship and changes in the fluctuation of the Capture Rate against number of test-set sites relationship under different training-set proportions.

(a) Power exponent b as a function of training-set proportion, with the red dashed line indicating the theoretical

value $b = 0.5$ and the blue band showing the 95% confidence interval; (b) Coefficient of determination R^2 as a function of training-set proportion, with the green band showing the 95% confidence interval; (c) Decay of the standard deviation of the Capture Rate across test-set size segments, with error bars indicating the standard error across training-set proportions.

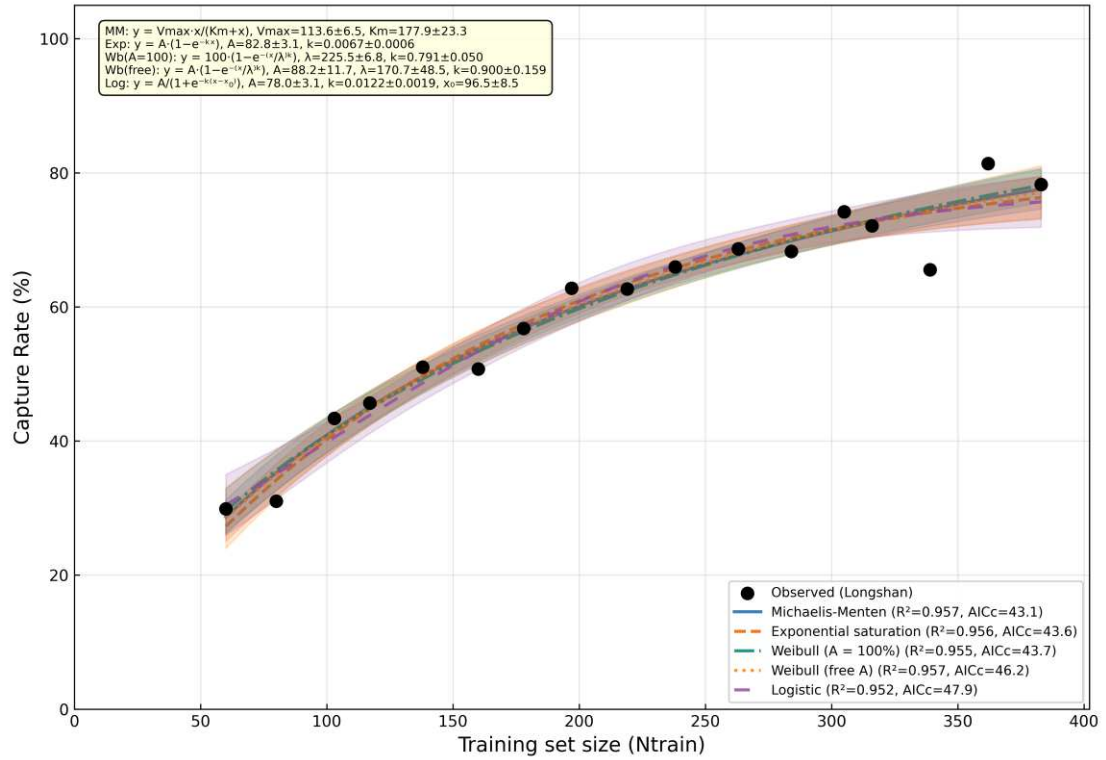


Figure 4. Relationship between the Capture Rate and the training-set size (N_{train}) under the condition of complementary partitioning.

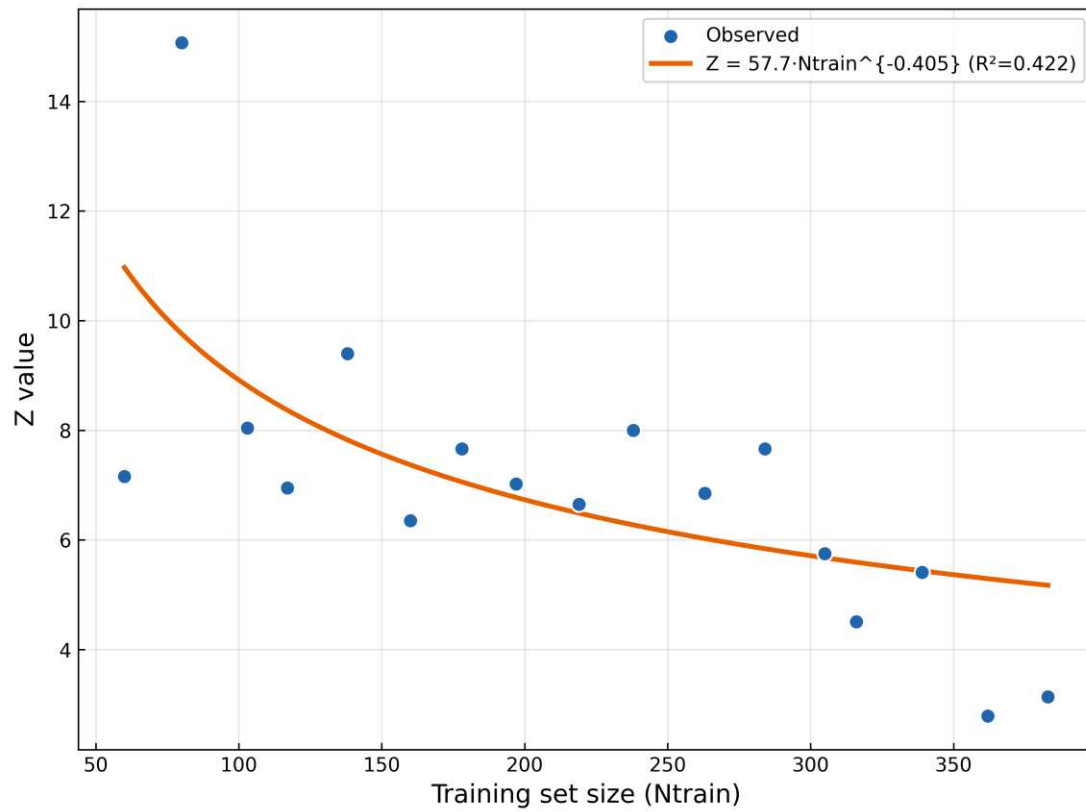


Figure 5. Relationship between the Z value and the training-set size (N_{train}) under the condition of complementary partitioning, indicating the presence of confounding by control-point size.

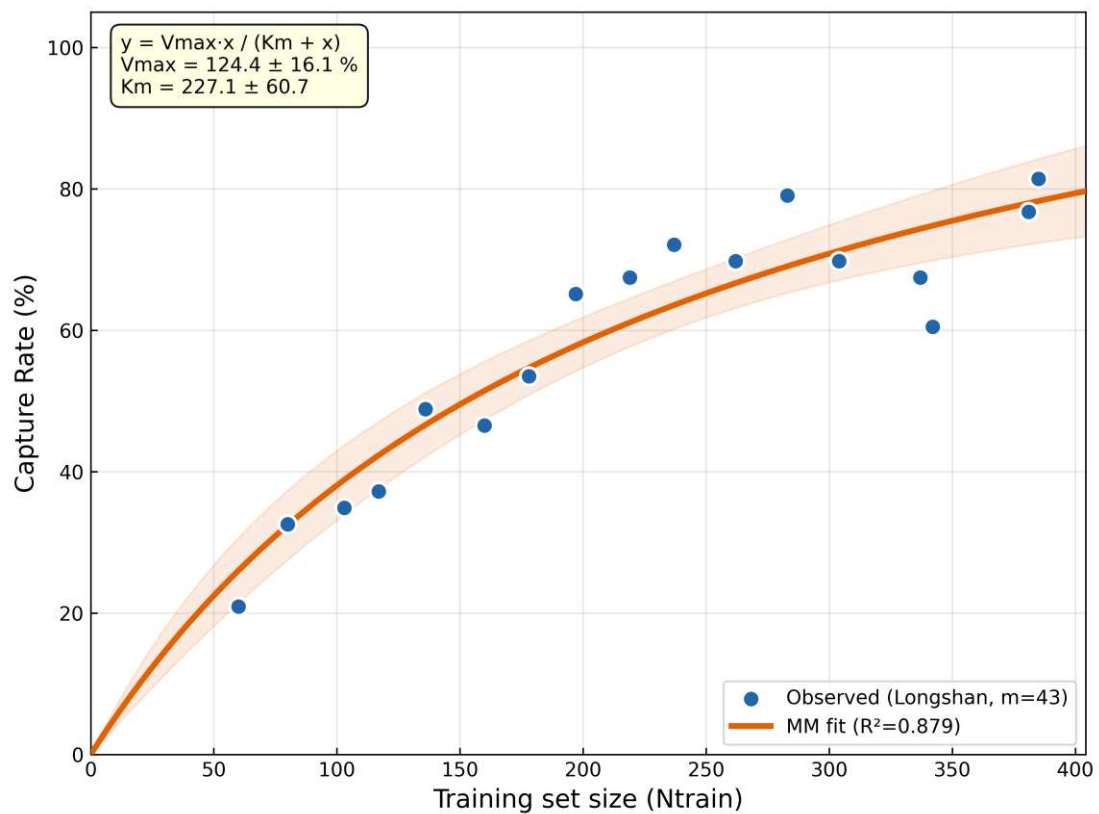


Figure 6. Relationship between the Capture Rate and the training-set size (N_{train}) when the control-point size is fixed at $m = 43$.

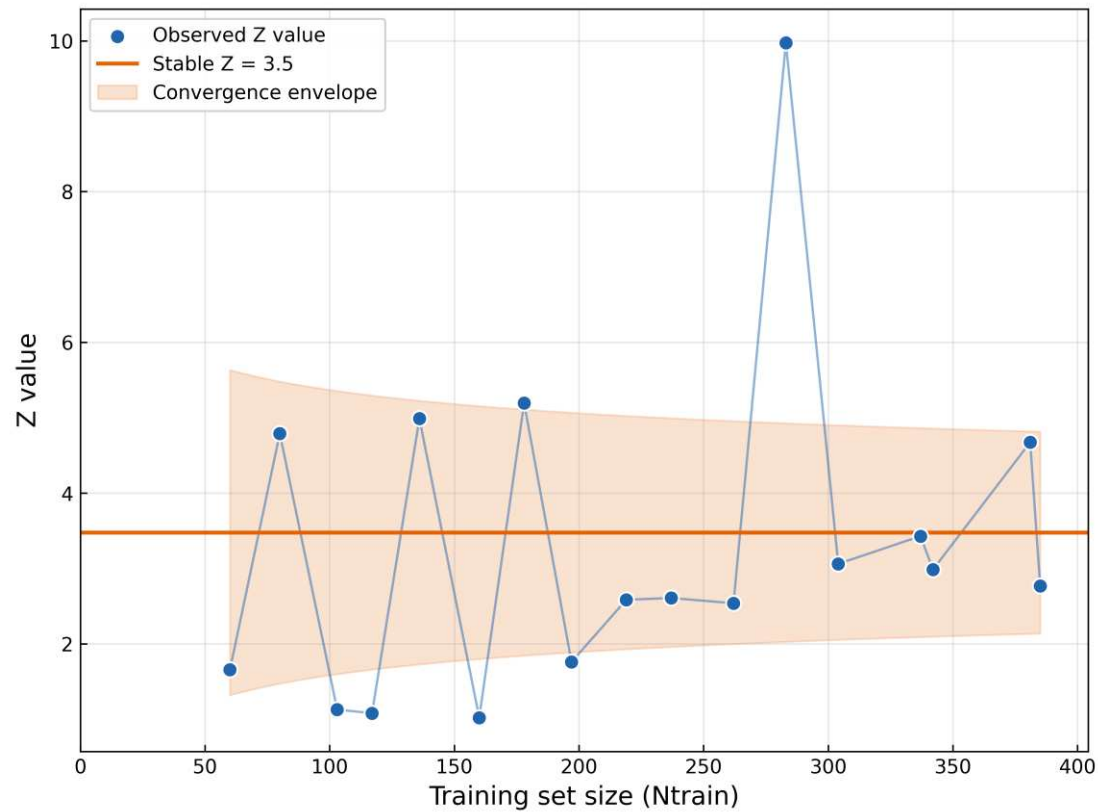


Figure 7. Relationship between the Z value and the training-set size (N_{train}) when the control-point size is fixed at $m = 43$.

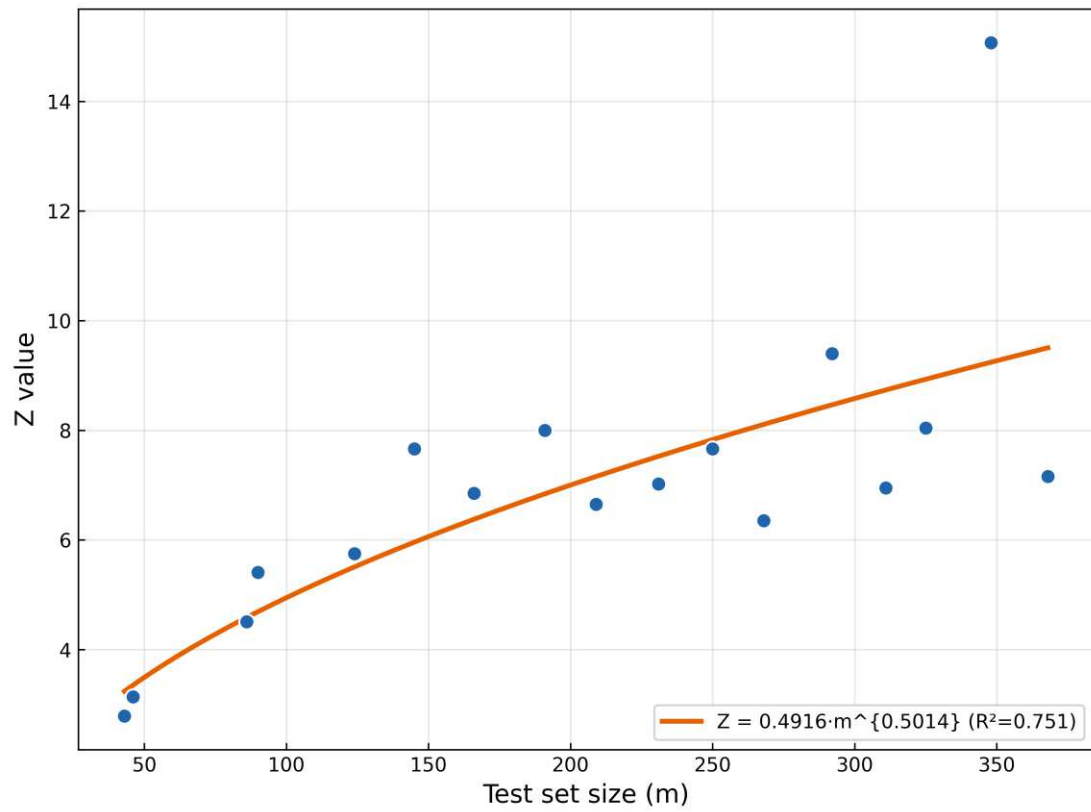


Figure 8. Relationship between the Z value and the number of test-set sites under the condition of complementary partitioning.

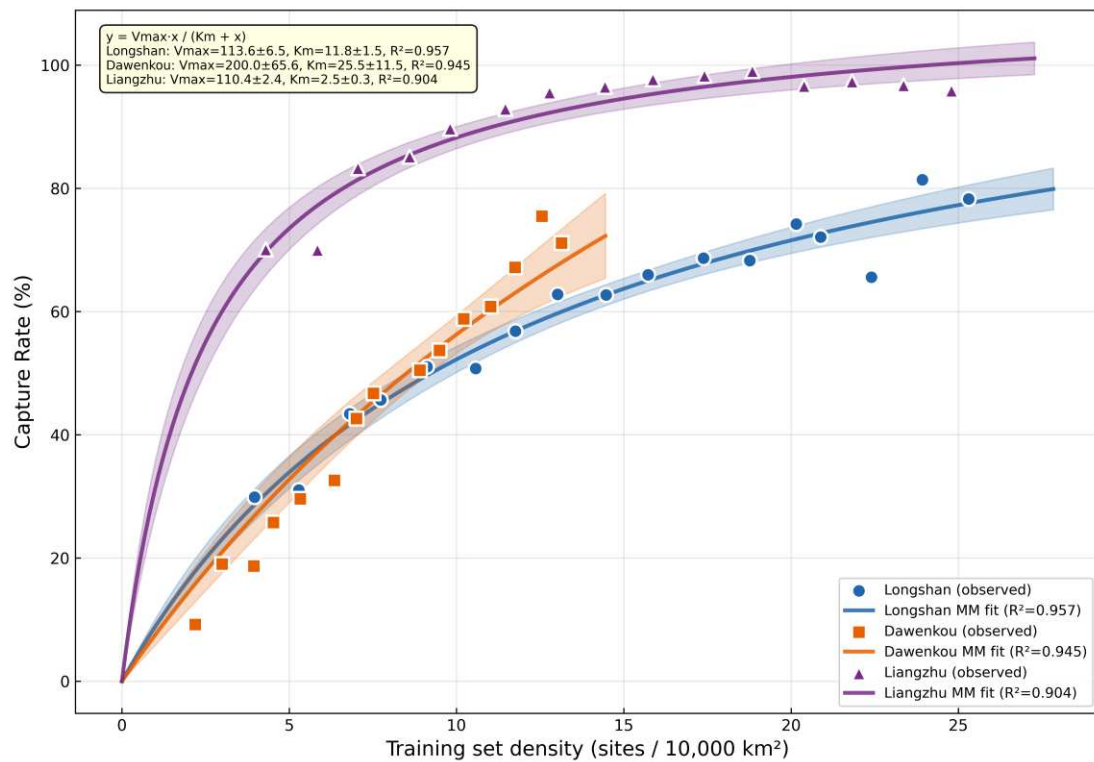


Figure 9. Relationship between Capture Rate and training-set density for three cultural periods.

Solid curves represent Michaelis-Menten fits ($y = V_{\max} \times x / (K_m + x)$), the best-performing model by AICc among five candidate saturation functions. See Table 3 for model comparison. All two-parameter saturation models yield nearly identical fitted curves ($\Delta AICc < 0.6$).

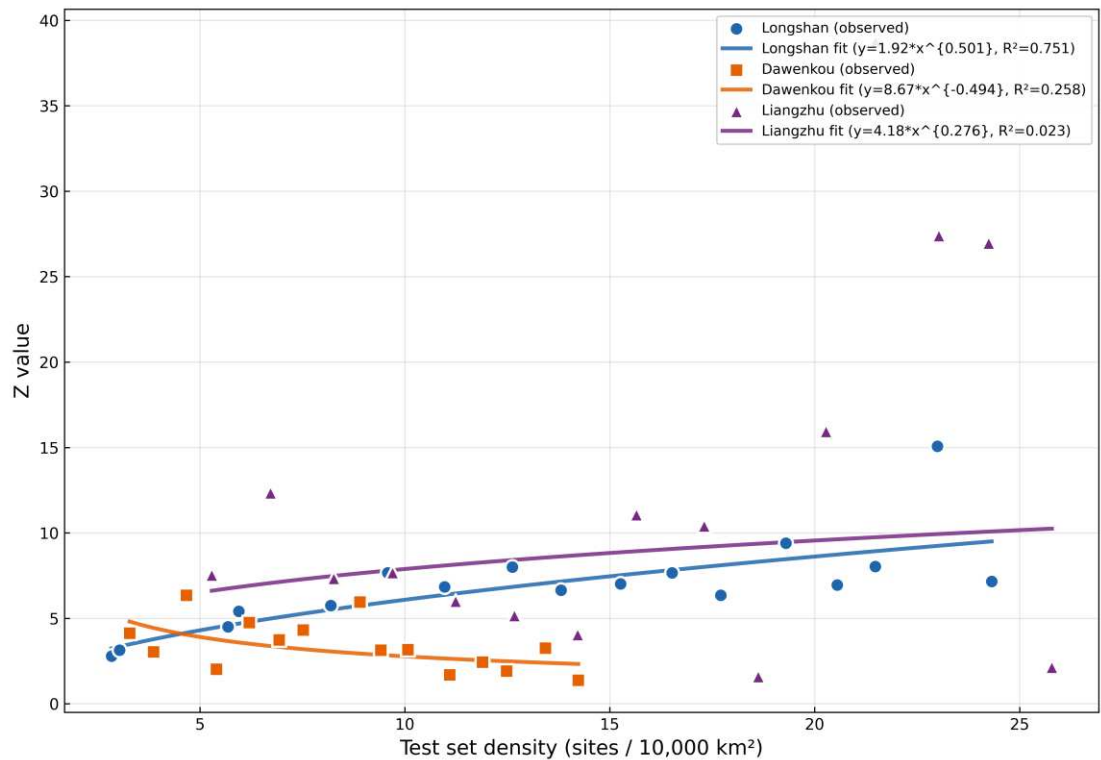
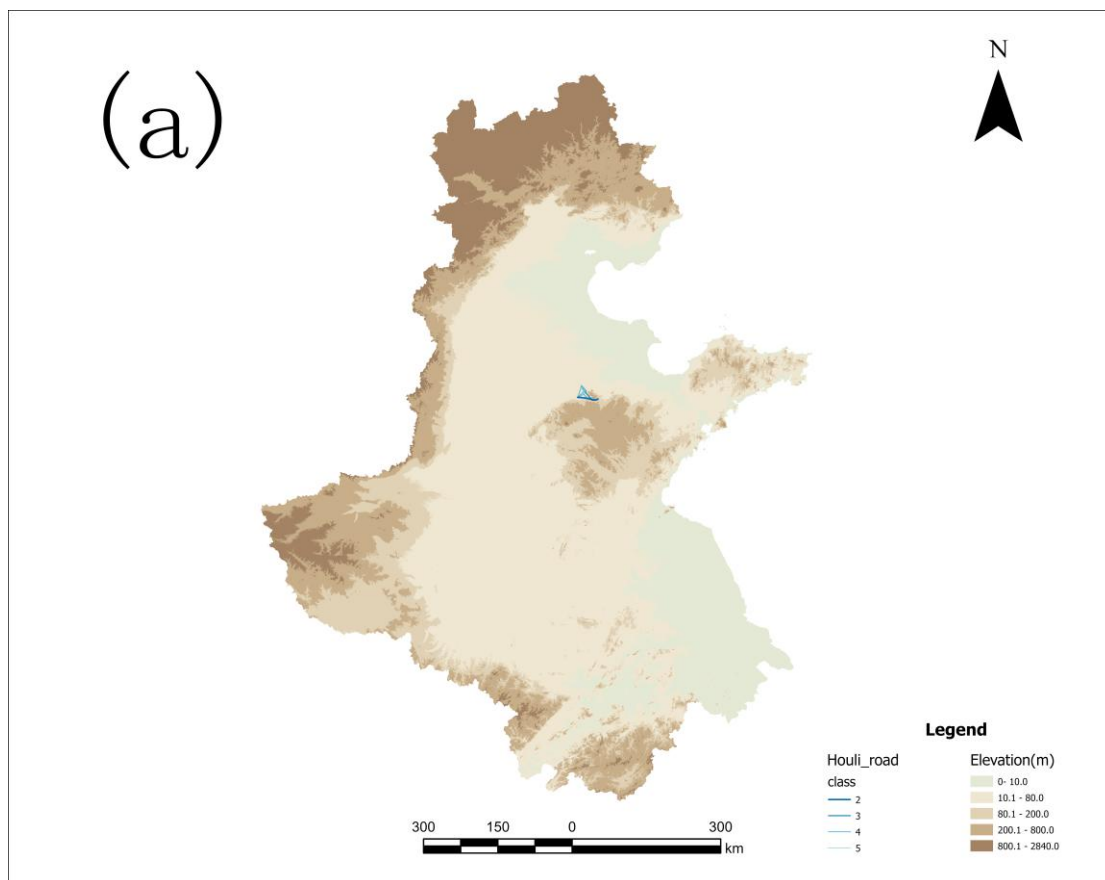
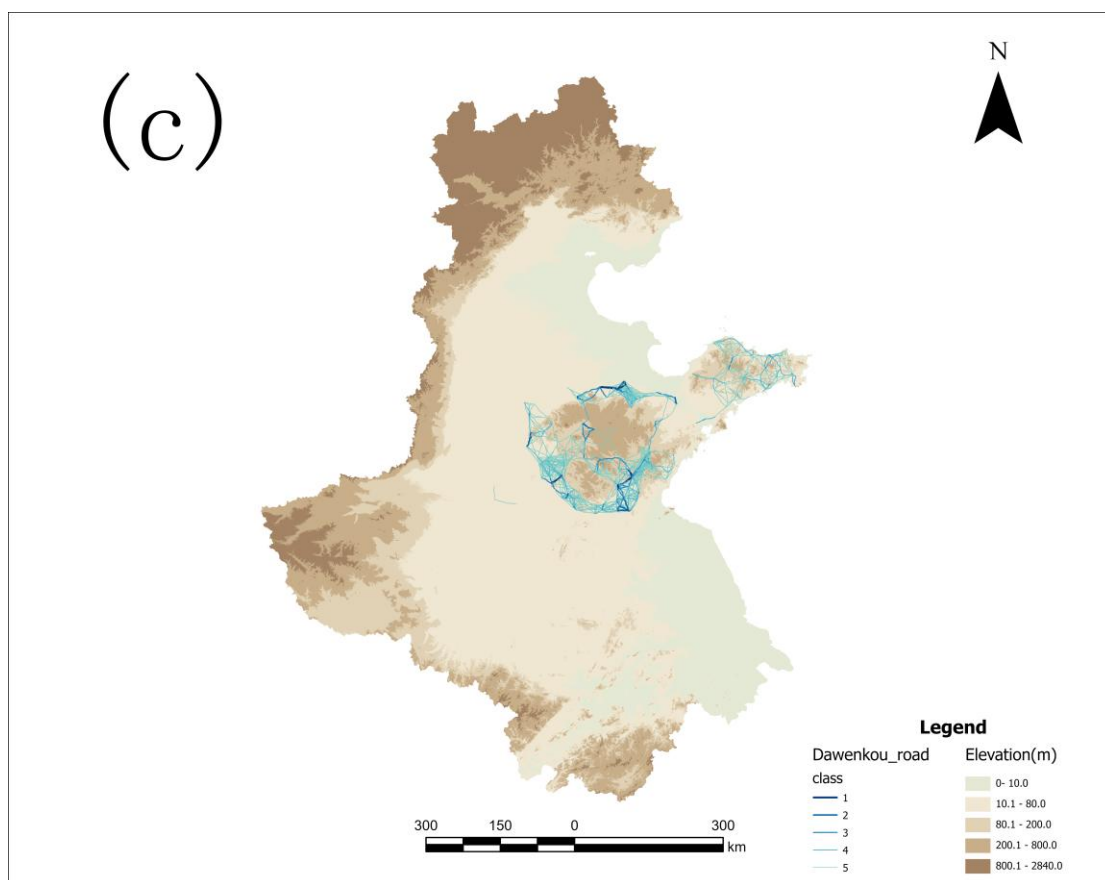
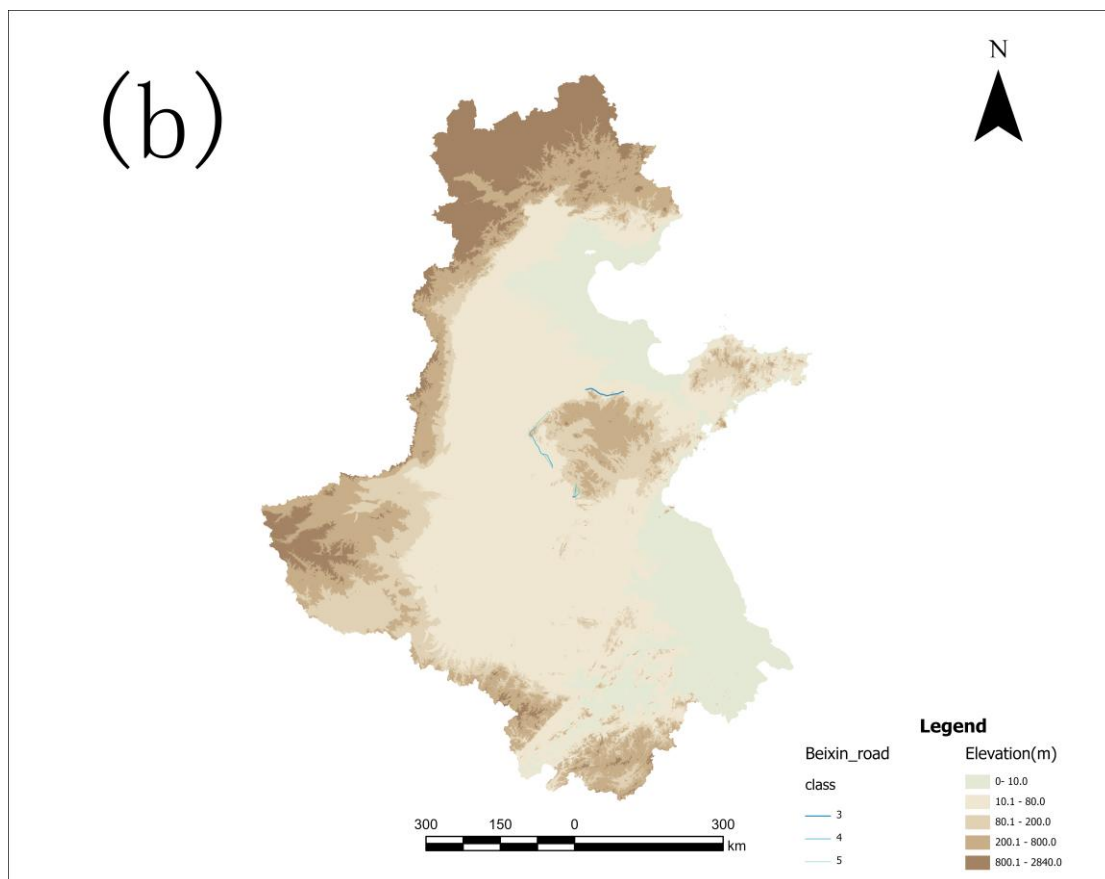


Figure 10. Fitted scatter plot of the Z value against test-set density relationship for three different cultural periods.





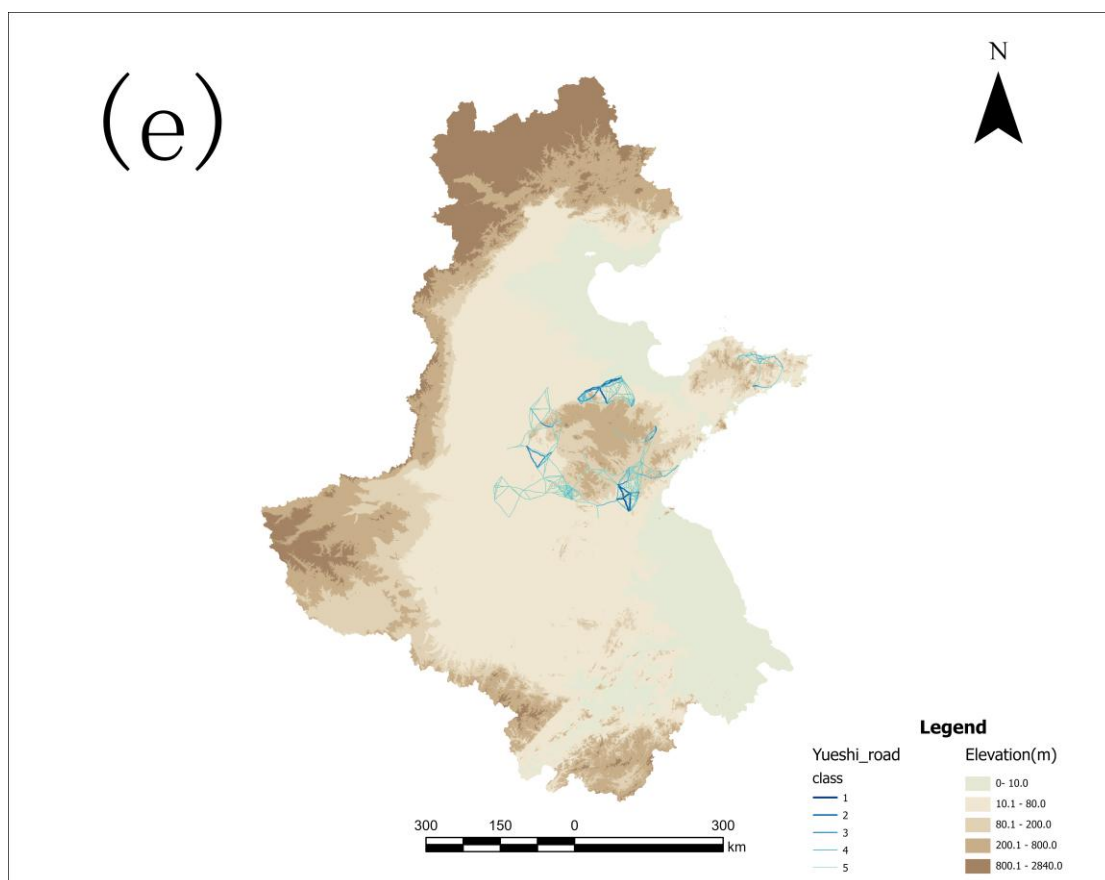
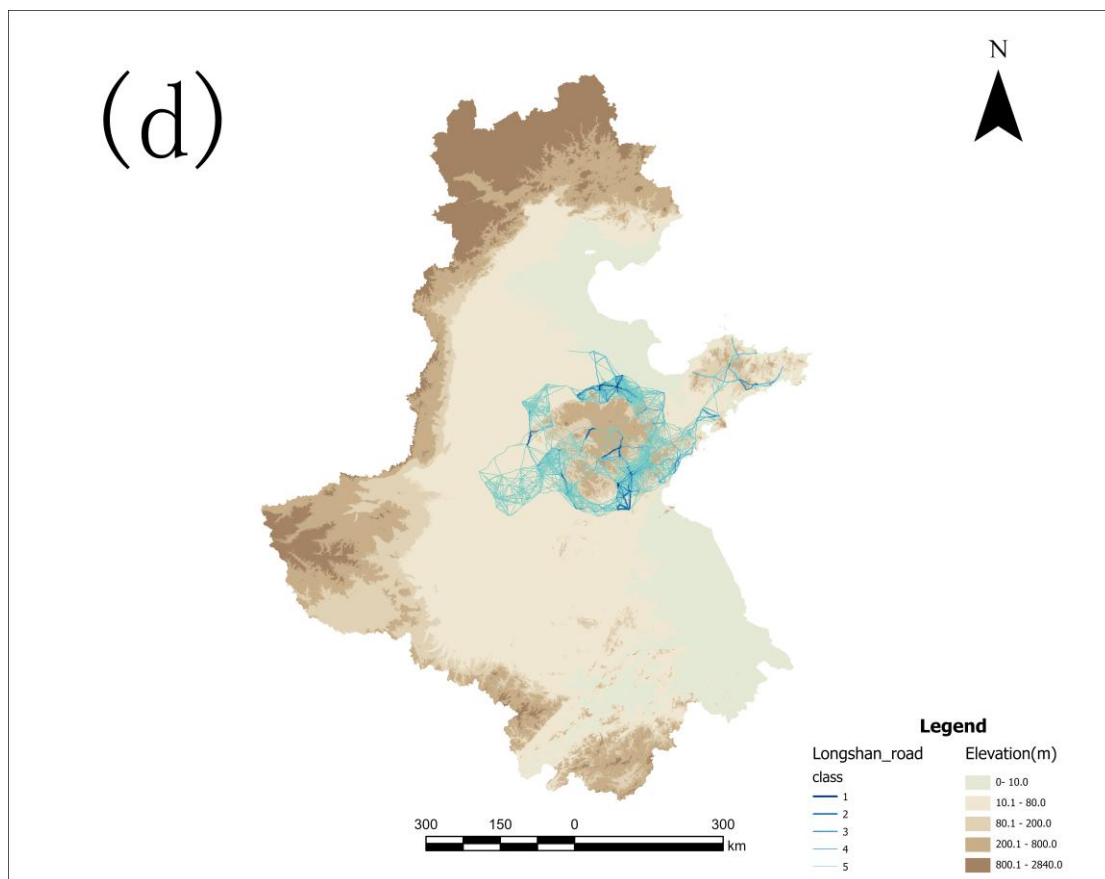


Figure 11. Relationship between road length and road hierarchy under the five-level road classification scheme. (a) Houli Culture; (b) Beixin Culture; (c) Dawenkou Culture; (d) Longshan Culture; (e) Yueshi Culture.

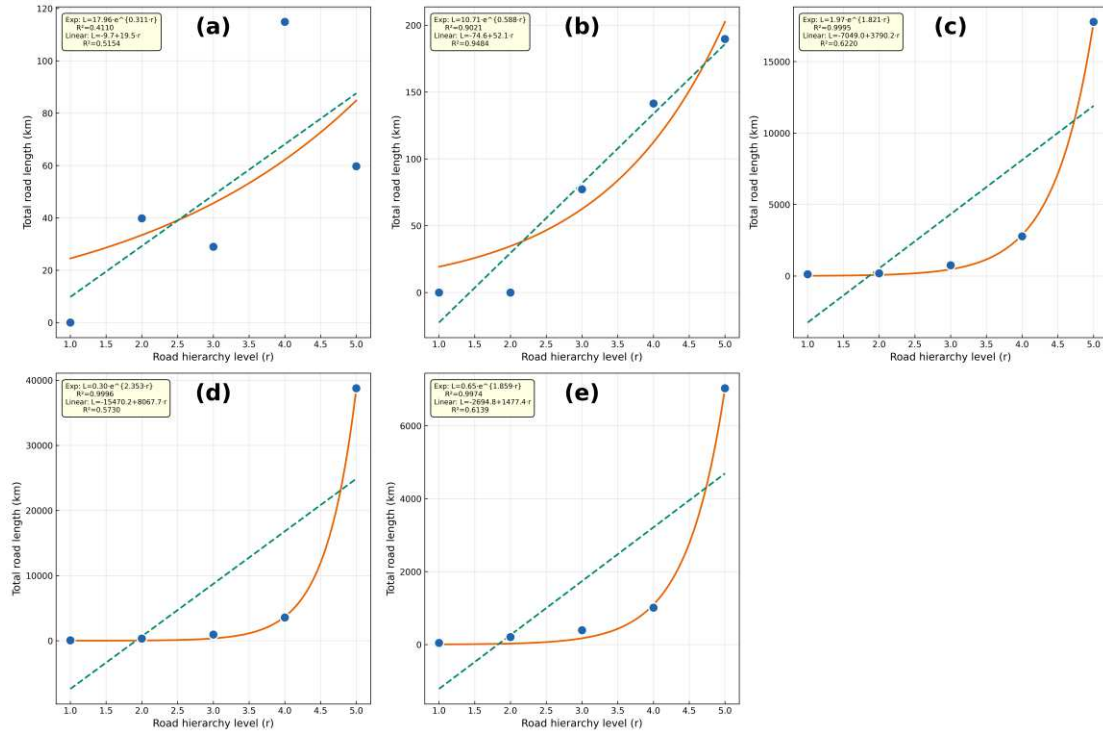


Figure 12. Variation in the coefficient of determination of the fitted functions with the number of road hierarchy levels.

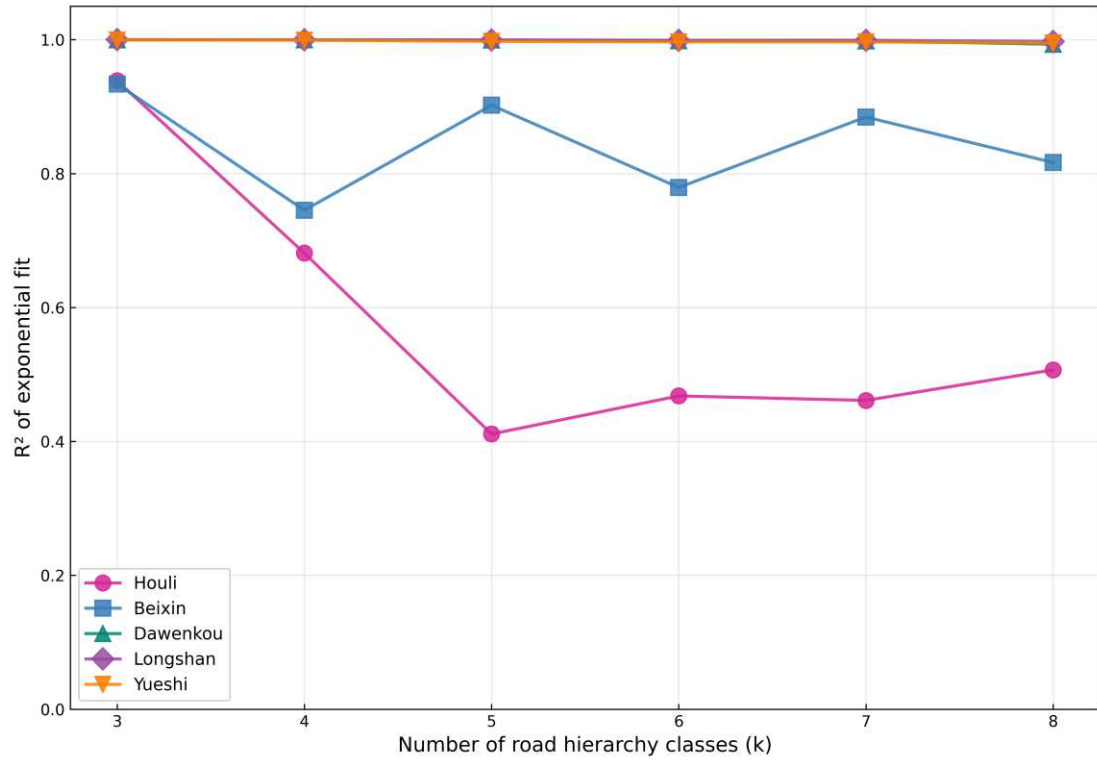


Figure 13. Five-level classification of road networks across the Neolithic cultural sequence in the Haidai region. (a) Houli Culture; (b) Beixin Culture; (c) Dawenkou Culture; (d) Longshan Culture; (e) Yueshi Culture.

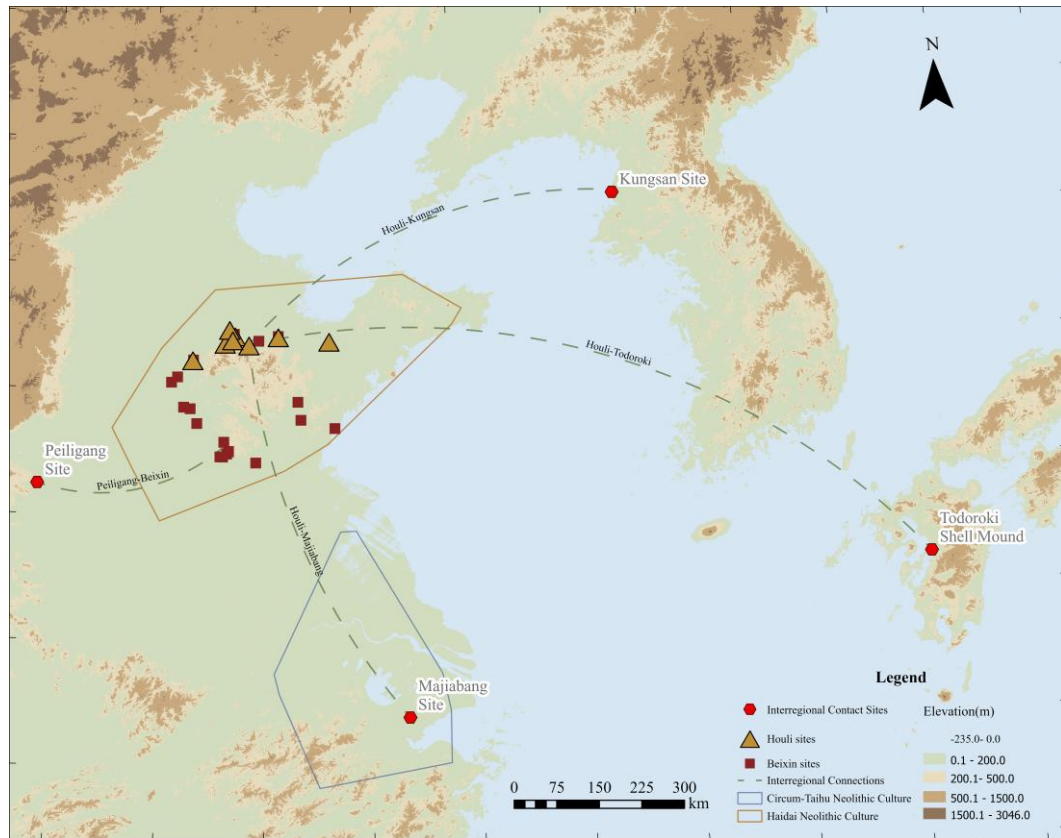


Figure 14. Study areas and materials. (a) Location of the Haidai region and the Taihu Basin in eastern China, with shaded relief showing topography; (b) distribution of Neolithic sites in the Haidai region, color-coded by cultural period (Houli, Beixin, Dawenkou, Longshan, Yueshi); (c) distribution of Liangzhu-period sites in the Taihu Basin.

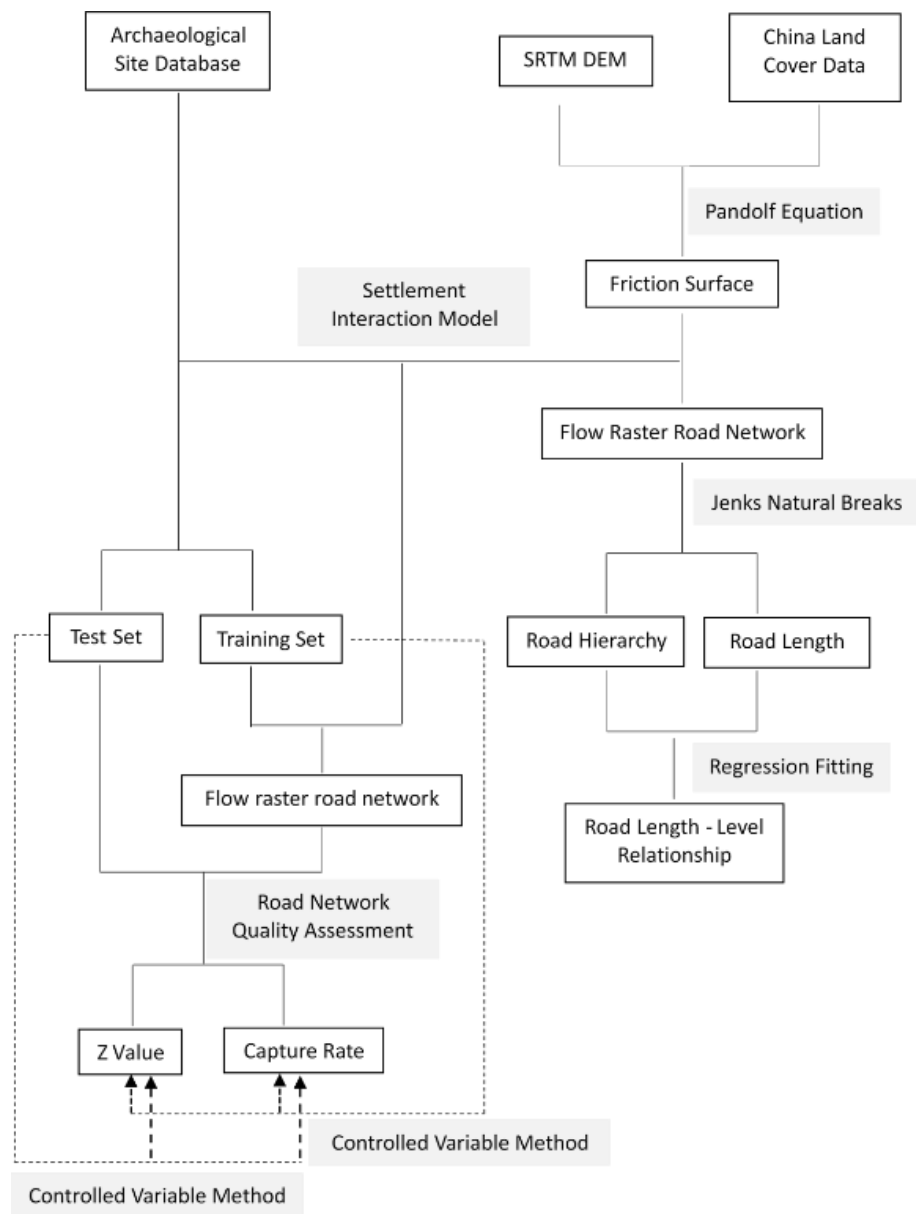


Figure 15. Schematic framework