

Review Article

The Rebirth of Ecological Study

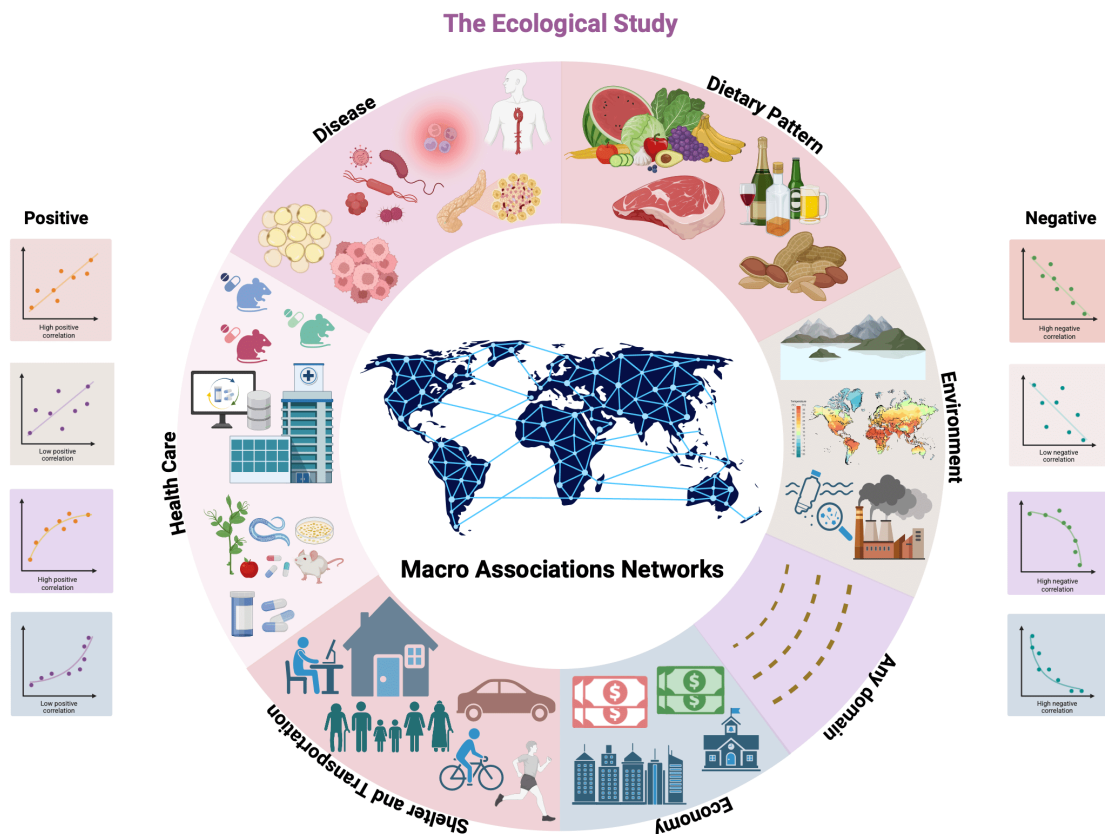
Hongyue Ma^{1,2}

1. School of Life Sciences, Westlake University, Hangzhou, China; 2. Haide College, Ocean University of China, China

Ecological study has long been criticized for its vulnerability to confounding and its reliance on simple group-level correlations. Here, this review introduces a structured methodological renewal that operationalizes ecological inference through macro association analysis (MA-analysis) and its formalized implementation, the Three Stars Global Macro Association-analysis (TSGMA) framework. Leveraging large-scale harmonized population-level datasets and scalable computation, MA-analysis reframes ecological study from isolated pairwise testing to staged, multi-variable, and temporally evaluated network construction. Through TSGMA, associations are systematically screened using predefined statistical thresholds, adjusted for representative macro-level confounders through structured selection procedures, and validated for temporal coherence across population units. These explicit criteria enable reproducible classification of associations into progressively stronger tiers and provide a scalable pathway for constructing one-star, two-star, and three-star MA networks. By situating ecological associations within a staged macro-analytic framework that integrates stability evaluation, structured confounder adjustment, and temporal assessment, MA-analysis shifts ecological inference from isolated pairwise interpretation toward macro-associative evaluation and network-based construction across variables and domains. This reframing enables ecological study to function as a structured macro-level mapping strategy, capable of organizing cross-domain determinants of health, sustainability, and societal systems within a coherent analytical architecture.

Corresponding author: Hongyue Ma, mahongyue@westlake.edu.cn

Graphical Abstract



1. Introduction

An ecological study is an observational study in which the units of analysis are populations or groups, rather than individuals^[1]. Historically, ecological studies played a foundational role in epidemiology, particularly when individual-level data were scarce or unobtainable^[2]. Landmark findings such as national dietary patterns and population-level disease burdens emerged from such studies, shaping early public health strategy^[3].

However, the method fell out of favor. In the late 20th century, limited data availability, lack of statistical standardization, and the absence of computational infrastructure undermined the reproducibility and perceived rigor of ecological studies^{[4][5]}. As evidence-based medicine advanced, so too did the emphasis on internal validity and individual-level inference. Increasing awareness of confounding and the ecological fallacy, whereby group-level associations are erroneously inferred to hold at the individual level, led to ecological studies being marginalized in the modern hierarchy of evidence^[6]. Therefore,

ecological studies, as a form of observational research, currently rank low on the evidence pyramid of evidence-based medicine, that is to say, they are considered among the weakest forms of evidence^{[7][8]}. Large-scale exploratory studies using population-level data became rare, and ecological designs were frequently dismissed as outdated or methodologically fragile.

At the core of this marginalization lies a structural limitation: traditional ecological analyses were typically restricted to isolated pairwise correlations with limited capacity to model multivariable context, temporal dynamics, or cross-domain interactions. As a result, observed associations were vulnerable to omitted-variable bias, misattribution of group-level effects, and instability across settings.

Yet the landscape has shifted dramatically. In the 21st century, the confluence of high-performance computation, algorithmic advances, and standardized global data infrastructure has ushered in a new era^{[9][10][11][12][13][14][15]}. Public datasets from hundreds of countries, covering health, diet, agriculture, policy, economics, and demography, are now openly available and increasingly harmonized across time and space. Paradoxically, even as data for population level analysis have flourished, ecological study has remained overlooked, constrained by historical stigma and by the absence of methodological modernization.

It's a missed opportunity. In the present era, however, the unprecedented availability of large-scale datasets on population and group, together with the conceptual advance of macro associations (MA), has given rise to an entirely new paradigm of ecological study^[16]. Rather than relying on isolated bivariate comparisons, MA-analysis embeds ecological associations within a staged, multi-layered macro-data architecture that integrates cross-domain variables across spatial and temporal scales. Through the Three Stars Global Macro Association-analysis (TSGMA), associations are screened, adjusted for measured confounders, and assessed for temporal consistency, thereby directly addressing the principal mechanisms underlying ecological fallacy and overinterpretation.

This paradigm leverages the simplest analytical approaches to interrogate the most complex systems: assembling diverse datasets defined at the population level, applying ecological study designs to explore broad patterns of potential associations, and systematically identifying novel linkages that may reshape prevailing research frameworks. By structuring and interconnecting previously fragmented datasets into MA networks, this approach enables the exploration of large-scale, high-dimensional associations and the discovery of emergent regularities governing human health, environment, and societal development. The robustness of this renewed ecological platform does not derive from eliminating confounding entirely, but from distributing inference across multiple variables, units, and temporal dimensions,

thereby increasing stability, reproducibility, and interpretive transparency relative to traditional pairwise ecological models. In doing so, the meaning of ecological study is elevated from narrow and often inaccurate pairwise correlation analyses to a comprehensive, scalable, and more reliable exploration of macro level associations across disciplines, and ultimately across domains. Such a shift marks the beginning of a new trajectory in which ecological study evolves into a critical tool for mapping the interconnected dynamics of human survival and planetary sustainability.

2. The rise, decline, and rebirth of ecological study

Ecological study, defined by its use of population or group level units as the basis of analysis, emerged as a pragmatic response to data scarcity in early epidemiology (Figure 1)^{[1][4]}. During the 1940s and 1950s, when individual-level surveillance systems were rare or nonexistent, researchers turned to national aggregates to interrogate public health patterns^{[3][17][18]}. This era laid the theoretical groundwork for population-level thinking, with early contributions such as Frost's (1934) tuberculosis analyses and Kuller's proposal of "macro epidemiology" exemplifying the use of aggregate data to generate population-wide insights^{[19][20][21][22][23][24]}.

The method flourished during the 1960s to 1980s, finding broad application in nutrition, environmental health, chronic disease epidemiology, and social determinants of health^{[2][25][26]}. Studies linking dietary fat to cardiovascular mortality or air pollution to respiratory disease provided compelling, although tentative, evidence of macro level exposures^{[27][28][29]}. However, this period also marked the rise of conceptual critiques. The "ecological fallacy," formally articulated by Robinson in 1950, warned of the dangers of misattributing group-level correlations to individual-level mechanisms^[6]. This was followed by methodologically grounded concerns about uncontrolled confounding, effect modification, and the misinterpretation of spatial autocorrelation, challenges that were difficult to address given the computational and statistical tools available at the time^{[5][30]}.

By the 1990s and early 2000s, ecological study had undergone a clear decline in status^{[4][31][32]}. The ascendance of evidence-based medicine (EBM), bolstered by randomized trials and large cohort designs, codified a new hierarchy of evidence in which ecological designs were demoted to a very low tier^{[33][34][35]}. This marginalization was reinforced by the rise of meta-analytic techniques and causal inference frameworks that emphasized internal validity^[35]. Despite the growing availability of national and global statistics from agencies like WHO, FAO, and the World Bank, these were rarely leveraged systematically.

As a result, ecological studies, although theoretically promising, remained methodologically fragmented, inconsistently reported, and increasingly viewed as exploratory relics of a pre-modern data era.

The COVID-19 pandemic, however, disrupted this trajectory in meaningful ways. Unlike endemic or non-communicable diseases, the effective control of a large-scale pandemic necessitates the acquisition and synthesis of vast amounts of population-level data to monitor and mitigate transmission dynamics, a task for which the ecological study design is inherently well-suited. Faced with a rapidly evolving global crisis, policymakers and scientists turned to population-level data to generate real-time situational awareness^{[36][37][38]}. From early cross-country comparisons of mortality and intervention timing to global dashboards mapping testing and vaccination coverage, ecological logic, including aggregation, comparison, and temporal mapping, re-entered the epidemiological mainstream^{[39][40][41]}. While limitations of such analyses were widely acknowledged, COVID-19 underscored the utility of rapid, population-scale insight generation in the absence of granular cohort data^{[42][43]}. Moreover, it accelerated the institutional standardization of global health metrics, catalyzed open data infrastructures, and validated the importance of population-level reasoning in times of uncertainty.

Recent post-2020 ecological analyses increasingly reflect this transformation in practice and illustrate how the field is evolving toward a broader relational framework. For example, the World Health Organization (WHO) now release annually updated cross-sectional datasets covering global cancer incidence across numerous cancer types (Global Cancer Observatory, <https://gco.iarc.fr>). Building upon these harmonized global data, one analysis constructed a global relational network of cancer incidence in 2020, identifying structured patterns of co-distribution among multiple cancer types at the global scale^[44]. Rather than examining isolated cancer categories, this approach organized them within a unified relational framework, revealing cross-type regularities that are not apparent in pairwise comparisons.

A further extension linked a dietary component network to the cancer network, thereby constructing an inter-domain association structure between nutritional patterns and cancer incidence^[45]. This integrative step moved beyond conventional correlation testing toward structured cross-domain mapping. Although still preliminary, such analyses demonstrate how ecological study can transition from isolated group-level correlations to network-based relational inference. In this sense, the rebirth of ecological study is not merely theoretical; it is already emerging in contemporary global data practice.

Today, a convergence of factors, including standardized longitudinal data, widespread open access, and maturing computational methods, suggests that the conditions are in place for a renewed form of ecological inquiry. Yet the question remains: can ecological study evolve beyond its historical limitations to become a robust, reproducible, and scalable scientific framework? This is not only possible, but necessary. Through the introduction of macro association analysis (MA-analysis), the use of TSGMA, and the construction of MA networks, ecological study can be transformed from a loose design category into a rigorous engine of exploratory discovery, capable of scanning structured macro-data for hypothesis generation and policy prioritization^[16].

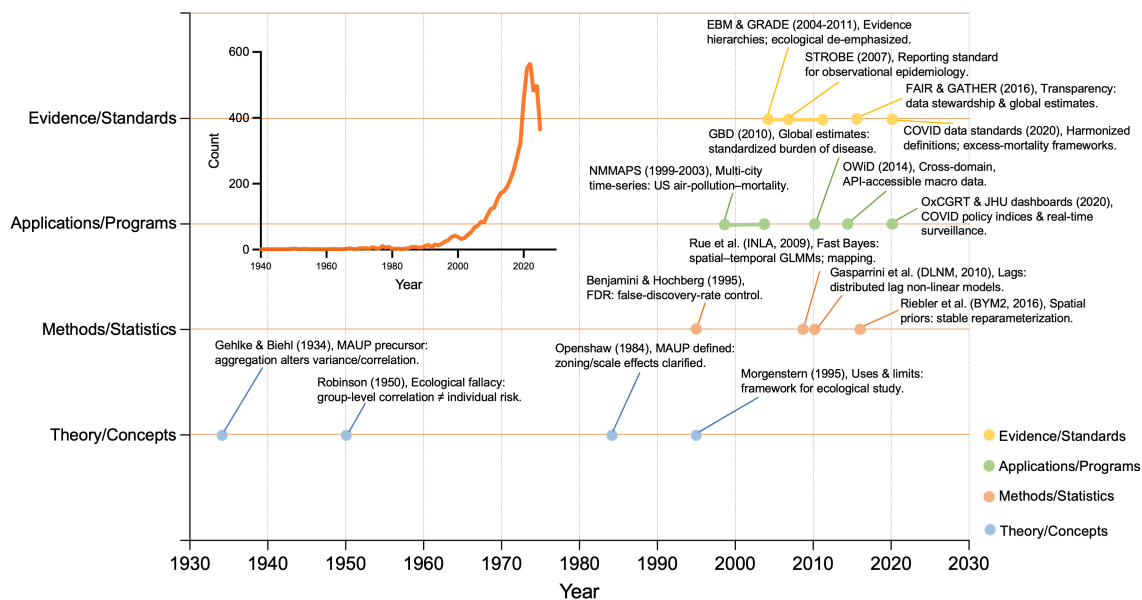


Figure 1. The development of ecological study. Publication counts are based on a rough PubMed search using “ecological study” as the title term. Since its introduction, the design has not generated a large volume of publications. Although the number of studies has shown a steady increase since the 1990s, the overall output has remained relatively modest. The emergence of COVID-19 prompted a surge of ecological studies that relied on population-level data to examine the pandemic, yet this expansion was largely confined to COVID-19 related applications rather than reflecting a broader methodological adoption. Following the peak of the pandemic, the number of publications has again shown a declining trend.

3. Unprecedented foundations: data, computation, and the new horizon for ecological study

3.1. Data explosion and infrastructural maturity

Data has always been the limiting substrate of ecological analysis. As a method relying on group-level statistics, ecological study is fundamentally constrained by the availability, coverage, granularity, and standardization of population-level data^{[1][4]}. For decades, such data was sparse, fragmented, and inconsistent across regions and time, which rendered ecological analysis narrow in scope and low in reproducibility^{[1][4]}.

The past decade, however, has witnessed a profound reversal. National statistical agencies, international organizations, and open-science platforms have converged toward standardized, high-resolution, longitudinal data systems. Massive, structured repositories, such as Our World in Data (<https://ourworldindata.org/>), FAOSTAT (<https://www.fao.org/statistics/en>), WHO Global Health Observatory (<https://www.who.int/data/gho>), and the World Bank Open Data (<https://databank.worldbank.org/databases>), now offer harmonized population-level variables across domains: health, diet, economics, climate, education, agriculture, governance, and beyond.

Moreover, these data are no longer scattered across PDF reports or proprietary silos. APIs, bulk CSV downloads, and machine-readable metadata have enabled automated, reproducible ingestion pipelines at scale. For ecological study, this constitutes a foundational shift: what was once a data-starved method is now data-rich. Yet paradoxically, most of these databases are limited to relatively simple displays of their contents, with few offering built-in systematic analyses or dynamically updating their analytical frameworks as new data are added. Only a small fraction is routinely analyzed, and even then, typically in domain-specific silos. The potential for cross-domain, cross-national, time-aware macro analysis remains largely untapped. This abundance of data alone does not guarantee analytic insight, but it does make ecological study technically feasible, globally scalable, and potentially transformative if supported by appropriate analytical frameworks. The previously described cancer relational network and the subsequent linkage between cancer incidence and dietary component networks further illustrate this point. Isolated analyses are limited to point-to-point associations and cannot reveal the broader structural interdependencies embedded within large databases. In the absence of systematic analytical integration, substantial relational information remains unobserved despite being readily extractable

through relatively simple computational procedures. In this context, network construction allows these datasets to be reorganized into more comprehensive relational structures, thereby enabling integrated analysis of cancer distributions and their cross-domain connections with dietary patterns.

3.2. Computational and algorithmic advances

Historically, ecological analysis depended on basic bivariate correlation and linear regression, often with minimal adjustment for confounding or structure^{[3][5][25][46]}. This simplicity was both a strength, as it permitted accessibility and interpretability, and a fatal weakness, as ecological studies were frequently dismissed as crude, over confounded, and methodologically fragile^{[2][4]}.

Today, that limitation is far less constraining. The computational landscape has been radically transformed. With the rise of cloud computing, GPU acceleration, distributed storage, and scalable statistical libraries, researchers can now process and analyze tens of thousands of variable pairs across countries and timepoints in minutes rather than months^{[47][48][49]}. Within the TSGMA framework, such scalability directly supports the one-star stage by enabling exhaustive pairwise screening across large variable sets, as well as repeated bootstrapped estimation to assess the stability of preliminary associations. This enables exhaustive screening, bootstrapped replication, rolling window validation, and lag-aware testing at population scale.

Crucially, advances in compositional data analysis, such as the centered log-ratio (CLR) transformation, allow variables that represent proportional components of a whole (for example, dietary shares or budget allocations) to be analyzed without generating spurious correlations due to their constrained sum structure^{[50][51]}. Together with robust estimation techniques, partial correlation networks, and causal graph theory, these methods provide tools to address many of the classical limitations of ecological study^{[52][53][54][55]}. In TSGMA, these techniques are most directly applied at the two-star stage, where partial correlation structures and multivariable adjustment are used to attenuate the influence of measured confounders and evaluate whether candidate associations persist under contextual control. Mixed models, hierarchical regression, and Bayesian priors allow structured control of multilevel confounding and regional heterogeneity^[56]. Ecological analysis, once methodologically static, is now technically modular, extensible, and statistically defensible.

Perhaps most transformative is the incorporation of Artificial Intelligence (AI)-driven workflows. Although current AI systems remain subject to limitations, including automation bias, model opacity, and

the potential amplification of underlying data biases^{[57][58][59]}, large language models and automated code synthesis now accelerate data wrangling, documentation, and sensitivity analysis^{[60][61][62][63]}. In practice, LLM-assisted workflows can standardize variable harmonization across datasets, automate reproducible pipeline generation, and facilitate systematic sensitivity testing across model specifications, thereby reducing implementation error in the construction of MA networks. AutoML platforms and explainable AI can prioritize association candidates, detect non-linear structures, and expose latent interaction patterns, while at the same time preserving interpretability constraints that are critical for ecological reasoning^{[53][64]}. Within the three-star stage, these tools can assist in identifying lag structures and non-linear temporal dependencies that may not be detectable through predefined parametric models alone, while still requiring human validation.

The combination of scalable computation and intelligent automation does not merely improve existing ecological methods; it operationalizes the staged logic of TSGMA. Specifically, computation enables breadth at the one-star level, structured adjustment at the two-star level, and temporal validation at the three-star level. In this context, the barrier to performing macro scale, high dimensional, and robust ecological analysis is no longer technical but conceptual. What is needed is a clear framework that governs discovery without overstating causality. This is the role assigned to TSGMA and MA networks, which are introduced in the next section.

4. The MA-analysis, MA networks, and ecological study

4.1. Conceptual foundations of MA-analysis and MA network architecture

Recently, a methodological innovation has been proposed that fundamentally reshapes the analytical paradigm of ecological study: the MA-analysis^[16]. MA-analysis is an analytical framework that incorporates as many cross-domain variables as feasible into structured correlation analysis, with the aim of constructing macro-associative networks (Figure 2). In essence, MA-analysis seeks to incorporate, to the greatest extent feasible, all measurable and computationally accessible macro-level variables relevant to a defined analytical scope into a unified relational network, thereby enabling association analysis from a comprehensive and system-wide perspective. By organizing associations across diverse domains within a unified relational architecture, MA-analysis enables the identification of both known and previously unrecognized patterns of interconnection. Through systematic examination of these macro-level structures, the framework supports retrospective understanding of developmental

trajectories as well as informed exploration of potential future dynamics across domains. This approach is centered on incorporating as many variables as possible in order to perform large-scale, cross-field exploratory analyses and to detect potential association networks at a macro scale (Figure 3). Its primary goal is to construct networks of associations within and across domains, to uncover hidden linkages, and to provide a broad framework that can guide future research and disciplinary development. Ultimately, MA-analysis aims to build MA networks and even pan-domain MA networks, connecting macro-level events across diverse dimensions of human society, the environment, and the economy, thereby enabling a deeper understanding of both societal progress and planetary evolution.

By complementing variables across dimensions, MA-analysis helps overcome one of the most persistent limitations of traditional ecological study: the distortion of results by unmeasured or dominant confounders in point-to-point analyses^{[4][5][30]}. Constructing single-domain MA networks transforms ecological study from the analysis of simple linear associations into multidimensional mapping within a domain. Extending this framework to cross-domain MA networks elevates ecological study from two-dimensional surfaces to multidimensional structures, fundamentally altering its scope and significance. For ecological study to achieve true conceptual breakthroughs, it must be integrated with the perspective of MA, thereby situating findings within higher-dimensional contexts.

The conceptual reach of MA-analysis should extend beyond population-level data. At a theoretical level, one may draw an analogy to other domains of large-scale relational modeling, for example by quantifying associations between the human genome and planetary ecology or between species level genomes and ecosystem development. By building such comprehensive relational networks, MA-analysis could ultimately support the creation of a “virtual Earth system”, analogous to the concept of a “virtual cell,” where diverse events are computationally integrated to reveal emergent laws of interaction^{[65][66][67][68]}. By analogy, just as genomics has enabled massively complex computations to quantify patterns across billions of bases, an Earth-scale MA framework could quantify virtually all recorded events, including human health, ecological, economic, and social data, and compute their interdependencies. The resulting networks would not only allow retrospective insights into human and planetary history, but also provide a foundation for forecasting future trajectories. Such a framework would represent a long-term theoretical aspiration rather than an immediately implementable system. If progressively developed, relational networks of this kind might enhance retrospective understanding of global patterns and potentially support structured forecasting.

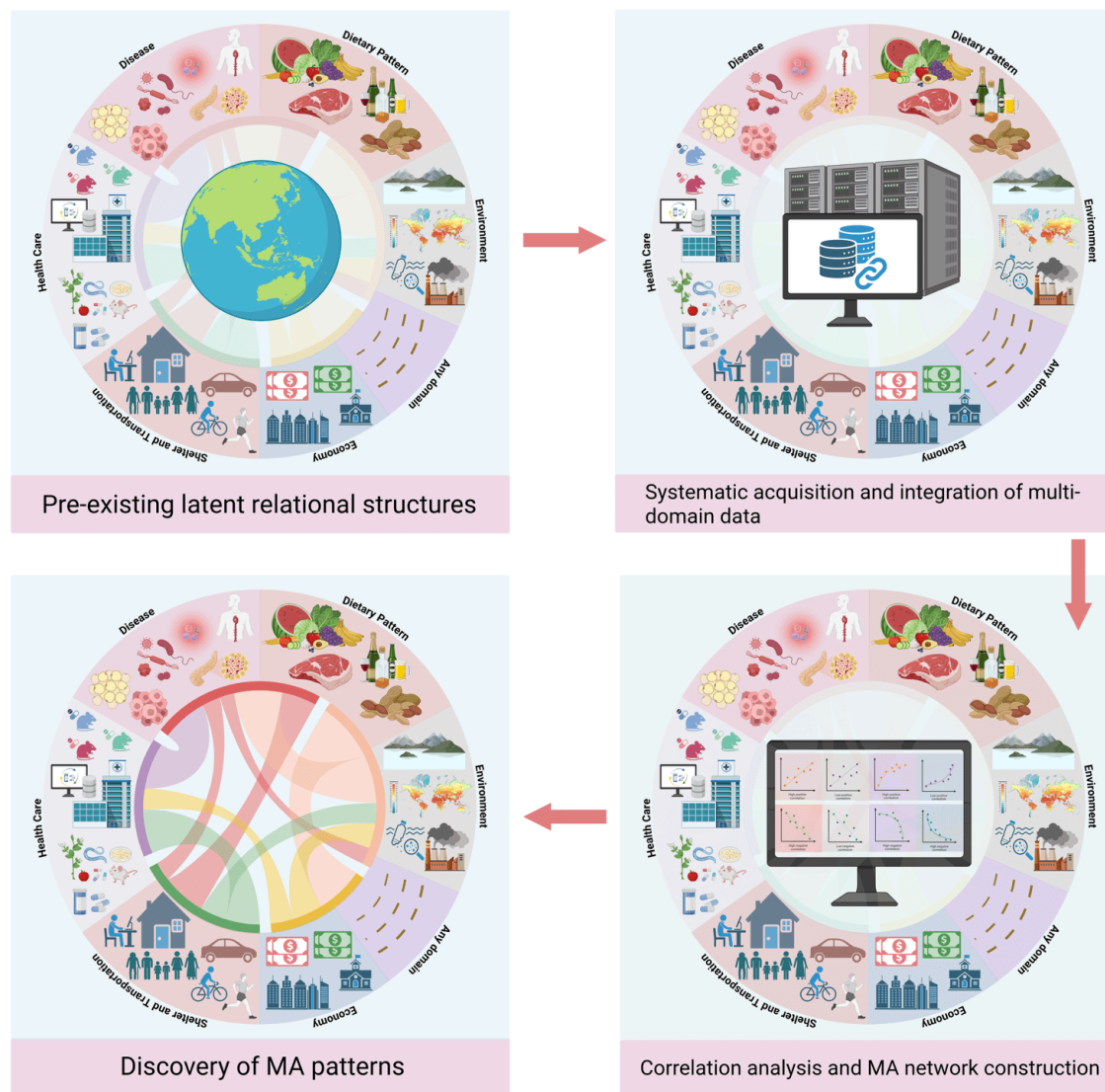


Figure 2. Conceptual workflow of MA-analysis. MA-analysis begins from pre-existing latent relational structures embedded across domains of health, environment, economy, and society. Through systematic acquisition and integration of harmonized multi-domain data, variables are organized within a unified macro-level analytical space. Correlation analysis is then performed to estimate inter-variable associations and construct MA networks. The resulting network structure enables the identification of macro-associative patterns, revealing structured regularities that may not be apparent in isolated analyses. Created with BioRender.

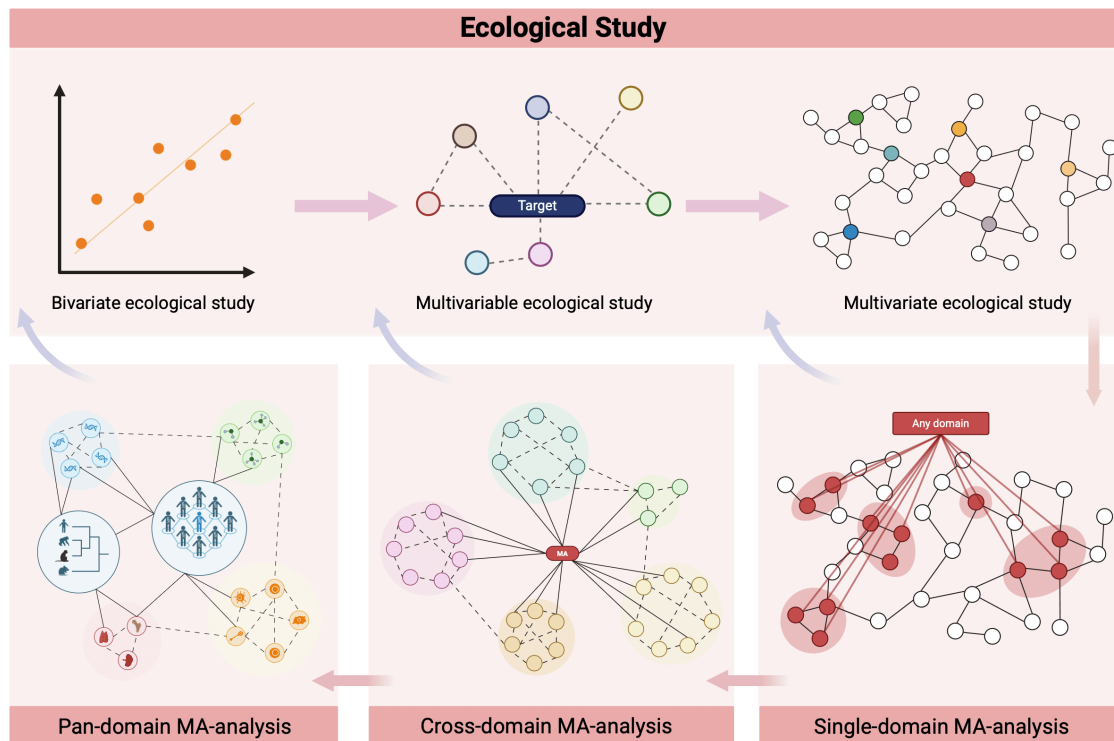


Figure 3. From traditional ecological study to MA-analysis. Conventional ecological studies, typically limited to pairwise associations, can be progressively extended to multivariable and multivariate analyses that capture multiple associations simultaneously, and further to the construction of small-scale local networks. In MA-analysis, these steps are organized within a structured macro-variable space in which associations are evaluated under predefined statistical thresholds and stability criteria rather than interpreted in isolation. When these localized analyses are expanded to encompass an entire disciplinary domain, they form the basis for single-domain MA networks. With further integration of harmonized variables across domains, such frameworks can evolve into cross-domain MA networks that span multiple fields, and ultimately into pan-domain MA networks that capture associations across broader layers of knowledge. At each stage of this progression, the constructed networks can be revisited under expanded contextual structures, allowing earlier associations to be re-examined through additional variable integration, confounder adjustment, and temporal evaluation, thereby strengthening interpretability without discarding prior results. Created with BioRender.

4.2. The operational framework and statistical implementation of TSGMA

Moreover, the development of the TSGMA framework provides ecological study with a structured and reproducible procedure for conducting MA-analysis. Each “star” represents a defined stage of evaluation

with explicit criteria for advancement, allowing the framework to function independently without reliance on unpublished materials^{[16][44]}.

One-star analysis begins with standardized pairwise association testing across variables defined at a comparable unit of analysis. In population-based applications, this typically involves harmonized population-level indicators. Correlation coefficients are calculated across population units at a specified spatial resolution. Statistical significance can be evaluated using two-sided tests or other appropriate inferential procedures, with adjustment for multiple comparisons when warranted. To ensure robustness, associations are further subjected to stability checks, such as resampling procedures or sensitivity analyses that assess whether results remain consistent when subsets of units are removed. Associations meeting predefined thresholds for statistical significance and stability are classified as one-star.

Two-star analysis introduces structured confounder adjustment. The transition from one-star to two-star follows a defined protocol in which a limited set of representative confounders is first identified. Confounder selection combines theoretical grounding with empirical evaluation. Variables that are widely recognized as macro-level contextual determinants, such as economic development, demographic composition, or healthcare capacity, are selected based on established literature and domain knowledge. In parallel, additional candidate variables are examined empirically and retained only if they demonstrate meaningful associations with both the exposure and the outcome, indicating potential confounding influence. To avoid over-adjustment and loss of interpretability, the number of retained confounders is deliberately restricted to those that capture broad contextual structure rather than every available variable. After confounder selection, partial correlation or multivariable regression analysis is performed. An association is promoted to two-star only if it remains statistically significant, maintains directional consistency, and does not exhibit excessive attenuation after adjustment. This ensures that the observed relationship cannot be fully explained by the selected macro-contextual variables.

Three-star analysis incorporates temporal validation. Using longitudinal or repeated cross-sectional data, temporal dynamics are evaluated within population units. Associations are examined to determine whether changes in one variable are consistently accompanied by changes in another over time, and whether plausible lead-lag patterns are observed. Promotion to three-star requires that the association demonstrate consistent temporal coherence across multiple units, not merely cross-sectional alignment at a single time point. In practice, this requires statistically supported temporal association within a predefined proportion of units and directional consistency across time windows. This stage strengthens

interpretability by reducing the likelihood that associations are driven by static structural differences alone.

In network construction, these three levels correspond to one-star, two-star, and three-star MA networks. For relatively modest sets of variables, such as those confined to a specific field with less variables, TSGMA provides accurate results and robust inference. However, as the dimensionality of MA-analysis expands to encompass entire single-domain or cross domains (for example, complete lifestyle profiles against all disease groups), applying full TSGMA at scale becomes computationally prohibitive. Thus, in practice, MA-analysis and TSGMA must be applied in a complementary and staged manner. When the primary aim is to determine accurate associations for a limited number of variables, TSGMA is appropriate. Conversely, when constructing large-scale MA networks, initial analyses may rely primarily on one-star TSGMA to build the preliminary network framework. Such a network, though exploratory, is meaningful at the macro level and inherently complementary across variables. Higher-order refinements can then be introduced gradually by incorporating two-star and three-star analyses, rather than attempting to complete full TSGMA prior to network construction (Figure 4).

However, the staged application of TSGMA requires explicit safeguards to avoid the risk that large MA networks become opaque structures built primarily on exploratory one-star correlations. To prevent methodological opacity, quality control and reproducibility are embedded at multiple levels of the MA workflow. First, one-star networks are constructed under predefined statistical thresholds, with documented criteria for significance, multiplicity control, and stability testing. All parameter settings, data preprocessing steps, and exclusion rules are fixed prior to network construction and fully recorded to ensure replicability. Second, the process of selective refinement from one-star to higher-tier validation is rule-based rather than discretionary. Subsets prioritized for two-star or three-star evaluation are determined through transparent criteria, such as effect size ranking, cross-unit consistency, or network centrality measures, rather than ad hoc selection. Third, iterative refinement is version-controlled. Each stage of network development generates a documented and shareable network state, allowing comparison across iterations and enabling independent replication, external validation, and open scrutiny. Fourth, network-level diagnostics are applied to detect structural instability, including sensitivity analyses under variable removal, threshold variation, or alternative confounder specifications. Associations that are highly sensitive to minor analytical perturbations are flagged rather than treated as stable links. In this framework, scalability does not imply relaxation of standards. Instead, breadth at the exploratory stage is balanced by progressive validation and explicit documentation. The iterative

expansion of MA networks is therefore designed to enhance transparency and reproducibility, rather than to obscure analytical uncertainty.

The resulting preliminary MA networks provide a foundational scaffold encompassing all variables. Any subset of this network can subsequently be extracted for more targeted TSGMA or for domain-specific, fine-grained investigations. In this way, MA-analysis and TSGMA together offer a scalable, iterative pathway: from broad exploratory mapping of MA to progressively refined, confounder-adjusted, and temporally validated networks.

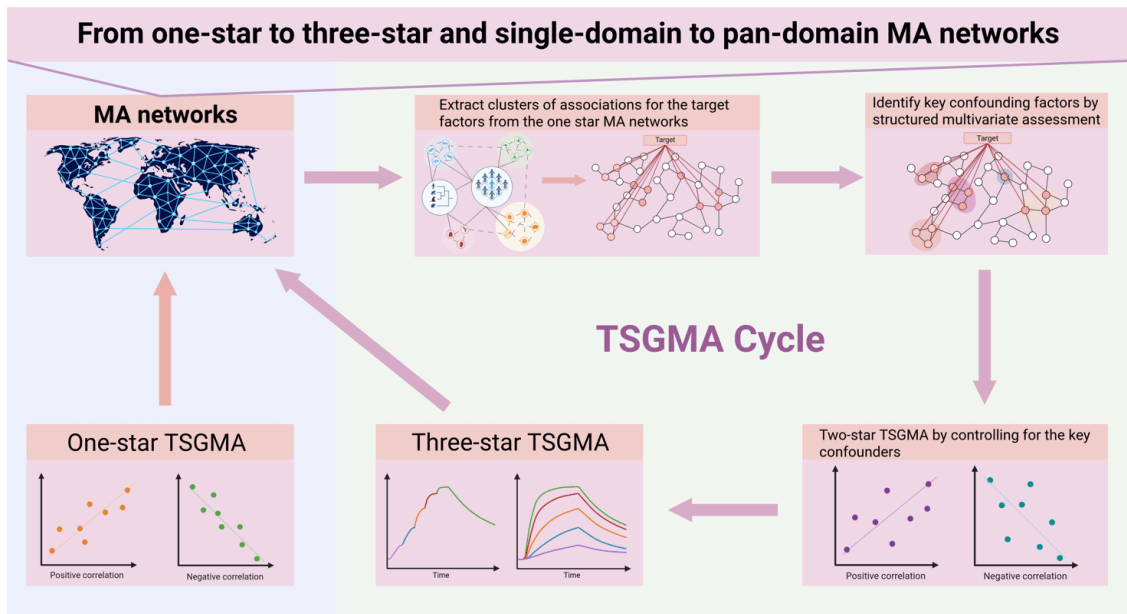


Figure 4. Construction and iterative refinement of MA networks. Using one-star TSGMA under predefined statistical significance and stability criteria, MA networks can be constructed across scales ranging from single-domain to pan-domain. These initial networks provide a comprehensive structural scaffold across variables. Once established, either specific target variables or the full network structure can be extracted for subsequent evaluation. After extracting a subnetwork centered on a target variable, the contextual structure of that subnetwork is examined to identify representative confounding factors influencing its associations. Confounder identification is conducted through a structured combination of domain-informed selection and empirical multivariate assessment, such as correlation structure evaluation, dimensionality reduction techniques including principal component analysis, or other diagnostics suited to the data architecture. Only a parsimonious subset of influential contextual variables is retained to preserve interpretability and prevent over-adjustment. Two-star TSGMA is then applied with structured confounder adjustment. Further application of three-star TSGMA incorporates temporal validation using repeated observations to assess cross-unit consistency and directional coherence over time. Associations that satisfy these higher-tier criteria are subsequently reintegrated into the broader MA network. This iterative refinement process constitutes the MA cycle, in which extraction, validation, and reintegration occur under explicit evaluation rules and documented parameter settings. Through repeated cycles, MA networks progressively transition from exploratory one-star structures toward higher-confidence two- and three-star configurations, ultimately consolidating into a stable and reproducible framework of macro associations. Created with BioRender.

5. Challenges in ecological study and MA network construction

As outlined above, confounding remains one of the most intractable limitations of ecological study and is a key factor constraining its development. Even in the two-star and three-star tiers of TSGMA, confounders can only be partially controlled. As an example, drawn from prior applications of TSGMA, analyses examining dietary components and non-high-density lipoprotein cholesterol (non-HDL-C) levels provide a useful illustration. The inverse association between vegetable intake and non-HDL-C is well established in epidemiological research^{[69][70]}. However, when dietary composition variables were evaluated within the TSGMA framework, a positive association between vegetable intake and non-HDL-C levels persisted across one-star and two-star analyses^[16]. This finding likely reflects residual or unmeasured confounding rather than a direct biological relationship. In particular, collinearity among dietary components may play an important role, as populations with high vegetable consumption may also exhibit concurrent intake patterns of other foods associated with elevated non-HDL-C levels. Such structural interdependencies among dietary variables cannot be fully eliminated within the TSGMA framework alone. Notably, subsequent multivariable regression analyses conducted in a related study attenuated this association, supporting the interpretation that it was driven by compositional and contextual confounding rather than causal effect. In conventional epidemiology as well, complete control of confounders is rarely feasible; most studies adjust only for a limited set of general covariates such as body mass index, smoking, alcohol consumption, or educational level, while many latent confounders remain unknown^{[71][72][73][74]}. Although these standard variables may capture a large portion of confounding, they cannot fully eliminate bias^{[71][72]}. Nevertheless, as systematic one-star MA networks are constructed and iteratively refined through the MA cycle, confounding may be progressively attenuated as additional variables provide proxy adjustment, redundancy, and opportunities for cross-checking. Empirically, high-dimensional covariate strategies can improve confounding control compared with adjustment limited to a small, predefined set, and modern high-dimensional inference frameworks can further enhance robustness through orthogonalization and resampling-based stability checks^{[75][76][77]}. Accordingly, in large-scale MA settings, expanding the variable space may increase the stability and reproducibility of identified relationships, rather than guaranteeing universally higher accuracy.

As noted earlier, ecological study is also vulnerable to spurious associations driven by collinearity and other structural factors^{[4][5]}. Yet an exclusive focus on direct associations may obscure the value of these spurious links. Even apparent false associations can signal underlying, meaningful processes; they

represent connections between variables that, while not directly causal, may still point to latent common determinants. By aggregating both true and spurious associations into large-scale networks, events can be grouped into meaningful clusters and examined collectively. Further internal analysis can then reveal which linkages are artifacts and which reflect genuine causal relationships. Importantly, once a sufficiently large one-star MA network is assembled, it begins to exhibit a self-correcting property: through clustering and internal comparisons, the network itself can help identify the factors responsible for spurious associations, highlight candidate causal pathways, and iteratively refine its own structure.

Because the construction of MA networks involves enormous data volumes, computational demands, and disciplinary breadth, it constitutes a major undertaking requiring extensive collaboration across fields and substantial investment of human and financial resources. Yet advances in computer science, the growing availability of open data, and the inherent simplicity of MA-analysis make this endeavor feasible. With the support of artificial intelligence, both the labor and cost required for such large-scale integration can be markedly reduced. At the same time, decisions regarding how special variables are defined, quantified, and incorporated into these networks, as well as how molecular level data will eventually be linked with macro level data, require systematic planning and standardization. Population units can also be progressively refined: from national or regional aggregates down to subnational units or even villages, enabling subgroup and diversity analyses. Such finer-grained stratification increases the number of groups, the diversity of comparisons, and the potential for nuanced inference, but it also requires balancing subgroup size against the precision and stability of results.

Beyond technical feasibility, the development of MA networks requires careful attention to ethical and governance frameworks. Issues of data sovereignty, especially in cross-national integration, must be addressed to ensure that countries and communities retain appropriate control over how their data are aggregated and interpreted. Equitable access to data infrastructures and analytical capacity is equally important, as large-scale computational systems risk reinforcing existing global asymmetries in knowledge production. Transparency in model construction, documentation of variable definitions, and reproducibility standards are essential to maintain scientific accountability. Furthermore, as analyses move toward finer-grained population units, safeguards for privacy protection and responsible data use become increasingly critical. Without such governance mechanisms, technical scalability alone cannot ensure legitimacy or societal trust.

Within the TSGMA framework itself, further challenges arise. In two-star analyses, the identification of confounders must be precise. In the context of large MA networks, there will inevitably be numerous

correlated potential confounders. Simply including all of them indiscriminately would obscure the original associations^{[16][78][79][80][81]}. Instead, it is necessary to identify the small subset of variables that exert the greatest influence and account for the majority of confounding. Some variables may display atypical patterns, appearing uncorrelated under one-star analysis but becoming associated after proper adjustment, which highlights the need to reconsider how such variables are analyzed and incorporated. Without this nuance, traditional TSGMA could overlook variables that are intrinsically related to the outcomes of interest. In three-star analyses, additional complexity arises from time-series effects, unpredictable fluctuations, and subtle lagged responses. Because each population unit requires its own time-series computation, the computational burden can be substantial. Developing rapid and accurate time series matching methods, potentially supported by artificial intelligence, will therefore be essential for scaling TSGMA to the three-star level.

In addition to the expansion in data availability noted in Section 3.1, important limitations remain regarding the scope, comparability, and quality of external data platforms. First, a trade-off often exists between breadth and depth. For example, databases such as Our World in Data cover a wide range of thematic domains, yet within each domain the number of variables may be limited and the granularity of measurement relatively coarse. Some indicators are available only for a small subset of countries or years, resulting in limited sample coverage and reduced statistical reliability. Second, the definition and operationalization of variables are critical. In MA-analysis, each event must be precisely defined and quantified using valid and comparable indicators. Ambiguity in variable definition, inconsistent measurement standards, or reliance on proxy metrics can introduce structural distortion into network construction and downstream inference. Third, data accuracy and completeness vary across regions. Systematic data collection at the population level remains challenging, and in some contexts, estimates are derived from modeling or extrapolation rather than direct reporting. As a result, uncertainty in underlying measurements may affect downstream analyses. Fourth, issues of standardization, accessibility, and data formatting influence usability. Harmonized variable definitions, transparent metadata, and machine-readable bulk download options are essential for reproducible large-scale analysis. In particular, under an MA-analytic framework, the ability to retrieve comprehensive datasets efficiently is critical. Finally, not all databases provide fully open access, and restrictions on data availability can limit cross-platform integration and independent validation. These considerations underscore that while global data infrastructures have expanded substantially, careful evaluation of

variable definition, coverage, comparability, and accessibility remains necessary when conducting large-scale ecological analyses.

6. Future directions of ecological study

When ecological study is reconsidered through the lens of MA networks, it can, for the first time, become truly “ecological.” Just as ecosystems are composed of populations yet operate not through isolated events but through interlinked processes that together sustain system dynamics, so too should ecological study move beyond the analysis of single associations. Instead, it should be conceived as the study of interconnected networks of events that collectively shape population- and planetary-level outcomes.

Only when the number of variables and associations reaches a sufficiently large scale can ecological study genuinely reflect the complexity implied by its name. The future of ecological study, therefore, lies in its transformation: from simple analyses of isolated associations, or its use as a subsidiary component within other study designs, toward systematic MA-analysis within specific domains; from there, to integrative frameworks spanning multiple domains; and ultimately, to the construction of pan-domain networks. At its core, the ultimate goal of MA-analysis and MA networks is to quantify the interconnections among all events in the world, from the molecular level to humanity and even the planetary scale, thereby revealing the full spectrum of linkages and constructing a comprehensive network of all events.

To translate this vision into concrete scientific practice, we are currently undertaking two near-term initiatives. First, we are constructing the first open-source, single-domain MA network in the field of cancer research. This network integrates cross-national data on cancer incidence, mortality, treatment indicators, and related contextual variables, and is being developed with an interactive visualization interface to enable transparent exploration of association structures. Second, building upon the Our World in Data platform, we are developing a cross-domain MA network that systematically maps associations across health, economic, environmental, and policy variables.

In subsequent phases, these two networks will be dynamically linked to facilitate structured exploration of potential cancer-related determinants across domains. Rather than aspiring immediately to a comprehensive “virtual Earth system,” this staged approach emphasizes reproducibility, openness, and incremental expansion. By beginning with domain-specific networks and progressively integrating them, MA-analysis can evolve through demonstrable, collaborative milestones.

Such networks hold the potential not only to illuminate the pathways by which human societies and ecosystems have developed, but also to provide a framework for anticipating future trajectories, thereby helping us to understand both where we come from and where we may be heading.

7. Conclusion

Ecological study, once marginalized as a method of limited precision, is now poised for a profound transformation. The convergence of standardized macro-level datasets, advances in computation, and the conceptual innovation of MA-analysis together provide the foundation for a new era. By scaling from one-star exploratory associations to multi-tiered, confounder-adjusted, and temporally validated networks, ecological study can evolve from simple pairwise comparisons into a robust framework for mapping the interconnected systems that shape human societies and the planet. The construction of single-domain or cross-domain MA networks, and ultimately pan-domain MA networks, offers not only a means of revisiting long-standing scientific questions with unprecedented depth but also a platform for identifying the key determinants of collective health, sustainability, and resilience. In practical terms, such structured association mapping could support public health surveillance by identifying emerging risk constellations, informing prioritization of preventive interventions, and guiding resource allocation across populations. In doing so, ecological study regains its relevance as a truly ecological science: one capable of integrating diverse layers of information, from molecular to planetary, into a coherent architecture for understanding both our past and our possible futures.

Statements and Declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

No potential competing interests to declare.

Author Contributions

Hongyue Ma performed all analyses and wrote the manuscript.

Resource and Data Availability

This study did not generate any new samples or unique reagents and does not report original code. The data presented in this paper are available from the lead contact upon reasonable request. Requests for additional information, resources, or reagents should be directed to Hongyue Ma at mahongyue@westlake.edu.cn and will be fulfilled where appropriate.

Use of Generative AI

During the preparation of this work the author used DeepSeek for proof reading and spell checking. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. ^{a, b, c, d}Morgenstern H (1995). "Ecologic Studies in Epidemiology: Concepts, Principles, and Methods." *Annu Rev Public Health*. 16:61–81. doi:[10.1146/annurev.pu.16.050195.000425](https://doi.org/10.1146/annurev.pu.16.050195.000425).
2. ^{a, b, c}Morgenstern H (1982). "Uses of Ecologic Analysis in Epidemiologic Research." *Am J Public Health*. 72(12):1336–1344. doi:[10.2105/ajph.72.12.1336](https://doi.org/10.2105/ajph.72.12.1336).
3. ^{a, b, c}Keys A (1957). "Diet and the Epidemiology of Coronary Heart Disease." *J Am Med Assoc*. 164(17):1912–1919. doi:[10.1001/jama.1957.62980170024007e](https://doi.org/10.1001/jama.1957.62980170024007e).
4. ^{a, b, c, d, e, f, g, h}Wakefield J (2008). "Ecologic Studies Revisited." *Annu Rev Public Health*. 29:75–90. doi:[10.1146/annurev.publhealth.29.020907.090821](https://doi.org/10.1146/annurev.publhealth.29.020907.090821).
5. ^{a, b, c, d, e}Greenland S, Morgenstern H (1989). "Ecological Bias, Confounding, and Effect Modification." *Int J Epidemiol*. 18(1):269–274. doi:[10.1093/ije/18.1.269](https://doi.org/10.1093/ije/18.1.269).
6. ^{a, b}Robinson WS (1950). "Ecological Correlations and the Behavior of Individuals." *Int J Epidemiol*. 15:351–357.
7. ^ΔFeigin VL, Howard G (2008). "The Importance of Epidemiological Studies Should Not Be Downplayed." *Stroke*. 39:1–2. doi:[10.1161/STROKEAHA.107.503250](https://doi.org/10.1161/STROKEAHA.107.503250).
8. ^ΔShakir S, Lane S (2025). "From a Pyramid to an Amphitheatre (Triangle to Circle): Embracing the Totality of Evidence for the Postmarketing Evaluation of the Safety of Therapeutic Interventions." *BMJ Open*. 15:e13538. doi:[10.1136/bmjopen-2025-103538](https://doi.org/10.1136/bmjopen-2025-103538).
9. ^ΔLeCun Y, Bengio Y, Hinton G (2015). "Deep Learning." *Nature*. 521:436–444. doi:[10.1038/nature14539](https://doi.org/10.1038/nature14539).

10. [△]Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al. (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Sci Data*. 3:160018. doi:[10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
11. [△]Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017). "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sens Environ*. 202:18–27. doi:[10.1016/j.rse.2017.06.031](https://doi.org/10.1016/j.rse.2017.06.031).
12. [△]Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. (2018). "The UK Biobank Resource with Deep Phenotyping and Genomic Data." *Nature*. 562:203–209. doi:[10.1038/s41586-018-0579-z](https://doi.org/10.1038/s41586-018-0579-z).
13. [△]Mathieu E, Ritchie H, Ortiz-Ospina E, Roser M, Hasell J, Appel C, Giattino C, Rod s-Guirao L (2021). "A Global Database of COVID-19 Vaccinations." *Nat Hum Behav*. 5:947–953. doi:[10.1038/s41562-021-01122-8](https://doi.org/10.1038/s41562-021-01122-8).
14. [△]Tatem AJ (2017). "WorldPop, Open Data for Spatial Demography." *Sci Data*. 4:170004. doi:[10.1038/sdata.2017.04](https://doi.org/10.1038/sdata.2017.04).
15. [△](2020). "Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019." *Lancet*. 396:1204–1222. doi:[10.1016/s0140-6736\(20\)30925-9](https://doi.org/10.1016/s0140-6736(20)30925-9).
16. [△]Ma H (2025). "TSGMA: Identification of Macro Associations from Global Data to Build Global MA Networks." *medRxiv*. 2025.2006.2009.25329257. doi:[10.1101/2025.06.09.25329257](https://doi.org/10.1101/2025.06.09.25329257).
17. [△]Langmuir AD (1963). "The Surveillance of Communicable Diseases of National Importance." *N Engl J Med*. 268:182–192. doi:[10.1056/nejm196301242680405](https://doi.org/10.1056/nejm196301242680405).
18. [△]Robinson W (2009). "Ecological Correlations and the Behavior of Individuals*." *Int J Epidemiol*. 38:337–341. doi:[10.1093/ije/dyn357](https://doi.org/10.1093/ije/dyn357).
19. [△]Frost WH (1995). "The Age Selection of Mortality from Tuberculosis in Successive Decades. 1939." *Am J Epidemiol*. 141:4–9; discussion 3. doi:[10.1093/oxfordjournals.aje.a117343](https://doi.org/10.1093/oxfordjournals.aje.a117343).
20. [△]Comstock GW (2001). "Cohort Analysis: W.H. Frost's Contributions to the Epidemiology of Tuberculosis and Chronic Disease." *Soz Praventivmed*. 46:7–12. doi:[10.1007/bf01318793](https://doi.org/10.1007/bf01318793).
21. [△]Rose G (1985). "Sick Individuals and Sick Populations." *Int J Epidemiol*. 14:32–38. doi:[10.1093/ije/14.1.32](https://doi.org/10.1093/ije/14.1.32).
22. [△]Omran AR (2005). "The Epidemiologic Transition: A Theory of the Epidemiology of Population Change. 1971." *Milbank Q*. 83:731–757. doi:[10.1111/j.1468-0009.2005.00398.x](https://doi.org/10.1111/j.1468-0009.2005.00398.x).
23. [△]Kuller LH (1999). "Circular Epidemiology." *Am J Epidemiol*. 150:897–903. doi:[10.1093/oxfordjournals.aje.a010097](https://doi.org/10.1093/oxfordjournals.aje.a010097).

24. [△](2005). "The Macro-Epidemiology of Influenza Vaccination in 56 Countries, 1997–2003." *Vaccine*. 23:5133–5143. doi:[10.1016/j.vaccine.2005.06.010](https://doi.org/10.1016/j.vaccine.2005.06.010).
25. [△][△]Armstrong B, Doll R (1975). "Environmental Factors and Cancer Incidence and Mortality in Different Countries, with Special Reference to Dietary Practices." *Int J Cancer*. 15:617–631. doi:[10.1002/ijc.2910150411](https://doi.org/10.1002/ijc.2910150411).
26. [△]Gray AM (1982). "Inequalities in Health. The Black Report: A Summary and Comment." *Int J Health Serv*. 12:349–380. doi:[10.2190/xxmm-jmqu-2a7y-hx1e](https://doi.org/10.2190/xxmm-jmqu-2a7y-hx1e).
27. [△]Lave LB, Seskin EP (1970). "Air Pollution and Human Health." *Science*. 169(3947):723–733. doi:[10.1126/science.169.3947.723](https://doi.org/10.1126/science.169.3947.723).
28. [△]Yerushalmy J, Hilleboe HE (1957). "Fat in the Diet and Mortality from Heart Disease; A Methodologic Note." *N Y State J Med*. 57:2343–2354.
29. [△]Pope CA 3rd (1991). "Respiratory Hospital Admissions Associated with PM10 Pollution in Utah, Salt Lake, and Cache Valleys." *Arch Environ Health*. 46:90–97. doi:[10.1080/00039896.1991.9937434](https://doi.org/10.1080/00039896.1991.9937434).
30. [△][△]Piantadosi S, Byar DP, Green SB (1988). "The Ecological Fallacy." *Am J Epidemiol*. 127:893–904. doi:[10.1093/oxfordjournals.aje.a114892](https://doi.org/10.1093/oxfordjournals.aje.a114892).
31. [△]Pekkanen J, Pearce N (2001). "Environmental Epidemiology: Challenges and Opportunities." *Environ Health Perspect*. 109:1–5. doi:[10.1289/ehp.011091](https://doi.org/10.1289/ehp.011091).
32. [△]Subramanian SV, Jones K, Kaddour A, Krieger N (2009). "Revisiting Robinson: The Perils of Individualistic and Ecologic Fallacy." *Int J Epidemiol*. 38:342–360; author reply 370–343. doi:[10.1093/ije/dyn359](https://doi.org/10.1093/ije/dyn359).
33. [△]Burns PB, Rohrich RJ, Chung KC (2011). "The Levels of Evidence and Their Role in Evidence-Based Medicine." *Plast Reconstr Surg*. 128:305–310. doi:[10.1097/PRS.0b013e318219c171](https://doi.org/10.1097/PRS.0b013e318219c171).
34. [△]Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH (2011). "GRADE Guidelines: 3. Rating the Quality of Evidence." *J Clin Epidemiol*. 64:401–406. doi:[10.1016/j.jclinepi.2010.07.015](https://doi.org/10.1016/j.jclinepi.2010.07.015).
35. [△][△]Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS (1996). "Evidence Based Medicine: What It Is and What It Isn't." *Bmj*. 312(7023):71–72. doi:[10.1136/bmj.312.7023.71](https://doi.org/10.1136/bmj.312.7023.71).
36. [△]Dong E, Du H, Gardner L (2020). "An Interactive Web-Based Dashboard to Track COVID-19 in Real Time." *Lancet Infect Dis*. 20:533–534. doi:[10.1016/s1473-3099\(20\)30120-1](https://doi.org/10.1016/s1473-3099(20)30120-1).
37. [△]Hasell J, Mathieu E, Beltekian D, Macdonald B, Giattino C, Ortiz-Ospina E, Roser M, Ritchie H (2020). "A Cross-Country Database of COVID-19 Testing." *Sci Data*. 7:345. doi:[10.1038/s41597-020-00688-8](https://doi.org/10.1038/s41597-020-00688-8).
38. [△]Mathieu E, Ritchie H, Ortiz-Ospina E, Roser M, Hasell J, Appel C, Giattino C, Rodés-Guirao L (2021). "A Global Database of COVID-19 Vaccinations." *Nat Hum Behav*. 5:947–953. doi:[10.1038/s41562-021-01122-8](https://doi.org/10.1038/s41562-021-01122-8).

39. [△]Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, Whittaker C, Zhu H, Berah T, Eaton JW, et al. (2020). "Estimating the Effects of Non-Pharmaceutical Interventions on COVID-19 in Europe." *Nature*. 584:257–261. doi:[10.1038/s41586-020-2405-7](https://doi.org/10.1038/s41586-020-2405-7).
40. [△]Hsiang S, Allen D, Annan-Phan S, Bell K, Bolliger I, Chong T, Druckenmiller H, Huang LY, Hultgren A, Krasovich E, et al. (2020). "The Effect of Large-Scale Anti-Contagion Policies on the COVID-19 Pandemic." *Nature*. 584:262–267. doi:[10.1038/s41586-020-2404-8](https://doi.org/10.1038/s41586-020-2404-8).
41. [△]Brauner JM, Mindermann S, Sharma M, Johnston D, Salvatier J, Gavenčiak T, Stephenson AB, Leech G, Altmann G, Mikulík V, et al. (2021). "Inferring the Effectiveness of Government Interventions against COVID-19." *Science*. 371. doi:[10.1126/science.abd9338](https://doi.org/10.1126/science.abd9338).
42. [△]Björk J, Modig K, Kahn F, Ahlbom A (2021). "Revival of Ecological Studies During the COVID-19 Pandemic." *Eur J Epidemiol*. 36:1225–1229. doi:[10.1007/s10654-021-00830-9](https://doi.org/10.1007/s10654-021-00830-9).
43. [△]Chen YC, Chow-In Ko P, Lee WC, Lien WC (2021). "Ecological Fallacy in COVID-19 Epidemiological Inference: Influenza Vaccination Rate as an Example." *J Formos Med Assoc*. 120:1655–1656. doi:[10.1016/j.jfma.2021.03.011](https://doi.org/10.1016/j.jfma.2021.03.011).
44. ^{a, b}Ma H, Qi X (2023). "Red Meat Consumption and Cancer Risk: A Systematic Analysis of Global Data." *Foods*. 12. doi:[10.3390/foods12224164](https://doi.org/10.3390/foods12224164).
45. [△]Ma H (2025). "Mapping Global Dietary Risk Factors for Cancer: A Structured Network-Based Analysis." *medRxiv*. 2025.2010.2011.25337797. doi:[10.1101/2025.10.11.25337797](https://doi.org/10.1101/2025.10.11.25337797).
46. [△]Dufault B, Klar N (2011). "The Quality of Modern Cross-Sectional Ecologic Studies: A Bibliometric Review." *Am J Epidemiol*. 174:1101–1107. doi:[10.1093/aje/kwr241](https://doi.org/10.1093/aje/kwr241).
47. [△]Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Moore R (2017). "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sens Environ*. 202.
48. [△]Meng X, Bradley J, Yavuz B, Sparks E, Talwalkar A (2015). "MLlib: Machine Learning in Apache Spark." *J Mach Learn Res*. 17:1235–1241.
49. [△]Melnik S, Gubarev A, Long JJ, Romer G, Shivakumar S, Tolton M, Vassilakis T (2010). "Dremel: Interactive Analysis of Web-Scale Datasets." *Commun ACM*. 3:114–123.
50. [△]Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Springer Netherlands.
51. [△]Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modeling and Analysis of Compositional Data*.
52. [△]Aitchison J (1982). "The Statistical Analysis of Compositional Data." *J R Stat Soc Series B Methodol*. 44:139–160. doi:[10.1111/j.2517-6161.1982.tb01195.x](https://doi.org/10.1111/j.2517-6161.1982.tb01195.x).

53. ^Δ^ΔFriedman J (2008). "Sparse Inverse Covariance Estimation with the Graphical Lasso." *Biostatistics*. 9:432–441.
54. ^ΔEpskamp S, Fried EI (2018). "A Tutorial on Regularized Partial Correlation Networks." *Psychol Methods*.
55. ^ΔEgozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). "Isometric Logratio Transformations for Compositional Data Analysis." *Math Geol*. 35:279–300.
56. ^ΔDiez-Roux AV (2000). "Multilevel Analysis in Public Health Research." *Annu Rev Public Health*. 21:171–192. doi:[10.1146/annurev.publhealth.21.1.171](https://doi.org/10.1146/annurev.publhealth.21.1.171).
57. ^ΔLyell D, Coiera E (2017). "Automation Bias and Verification Complexity: A Systematic Review." *J Am Med Inform Assoc*. 24:423–431. doi:[10.1093/jamia/ocw105](https://doi.org/10.1093/jamia/ocw105).
58. ^ΔRudin C (2019). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nat Mach Intell*. 1:206–215. doi:[10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x).
59. ^ΔMehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021). "A Survey on Bias and Fairness in Machine Learning." *ACM Comput Surv*. 54:Article 115. doi:[10.1145/3457607](https://doi.org/10.1145/3457607).
60. ^ΔLiu L, Hasegawa S, Sampat SK, Xenochristou M, Chen WP, Kato T, Kakibuchi T, Asai T (2024). "AutoDW: Automatic Data Wrangling Leveraging Large Language Models." *Association for Computing Machinery. Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*.
61. ^ΔDvivedi SS, Vijay V, Pujari SLR, Lodh S, Kumar D (2024). "A Comparative Analysis of Large Language Models for Code Documentation Generation." *Association for Computing Machinery. Proceedings of the 1st ACM International Conference on AI-Powered Software*.
62. ^ΔXu W, Luo G, Meng W, Zhai X, Zheng K, Wu J, Li Y, Xing A, Li J, Li Z, et al. (2025). "MRAgent: An LLM-Based Automated Agent for Causal Knowledge Discovery in Disease via Mendelian Randomization." *Brief Bioinform*. 26. doi:[10.1093/bib/bbaf140](https://doi.org/10.1093/bib/bbaf140).
63. ^ΔReason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B (2024). "Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models." *Pharmacoecon Open*. 8:205–220. doi:[10.1007/s41669-024-00476-9](https://doi.org/10.1007/s41669-024-00476-9).
64. ^ΔFriedman JH (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*. 29:1189–1232.
65. ^ΔLoew LM, Schaff JC (2001). "The Virtual Cell: A Software Environment for Computational Cell Biology." *Trends Biotechnol*. 19:401–406. doi:[10.1016/s0167-7799\(01\)01740-1](https://doi.org/10.1016/s0167-7799(01)01740-1).
66. ^ΔKarr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B Jr, Assad-Garcia N, Glass JI, Covert MW (2012). "A Whole-Cell Computational Model Predicts Phenotype from Genotype." *Cell*. 150:389–401. doi:[10.1016/j.cell.2012.08.033](https://doi.org/10.1016/j.cell.2012.08.033).

67. [△]Hazeleger W, Aerts JPA, Bauer P, Bierkens MFP, Camps-Valls G, Dekker MM, Doblas-Reyes FJ, Eyring V, Finkenauer C, Grundner A, et al. (2024). "Digital Twins of the Earth with and for Humans." *Communications Earth & Environment*. 5:463. doi:[10.1038/s43247-024-01626-x](#).
68. [△]Bauer P, Stevens B, Hazeleger W (2021). "A Digital Twin of Earth for the Green Transition." *Nature Climate Change*. 11:80–83. doi:[10.1038/s41558-021-00986-y](#).
69. [△]Dauchet L, Amouyel P, Hercberg S, Dallongeville J (2006). "Fruit and Vegetable Consumption and Risk of Coronary Heart Disease: A Meta-Analysis of Cohort Studies." *J Nutr*. 136:2588–2593. doi:[10.1093/jn/136.10.2588](#).
70. [△]Aune D, Giovannucci E, Boffetta P, Fadnes LT, Keum N, Norat T, Greenwood DC, Riboli E, Vatten LJ, Tonstad S (2017). "Fruit and Vegetable Intake and the Risk of Cardiovascular Disease, Total Cancer and All-Cause Mortality—a Systematic Review and Dose-Response Meta-Analysis of Prospective Studies." *Int J Epidemiol*. 46:1029–1056. doi:[10.1093/ije/dyw319](#).
71. [△]VanderWeele TJ, Ding P (2017). "Sensitivity Analysis in Observational Research: Introducing the E-Value." *Ann Intern Med*. 167:268–274. doi:[10.7326/m16-2607](#).
72. [△]Fewell Z, Davey Smith G, Sterne JA (2007). "The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study." *Am J Epidemiol*. 166:646–655. doi:[10.1093/aje/kwm165](#).
73. [△]Nørgaard M, Ehrenstein V, Vandenbroucke JP (2017). "Confounding in Observational Studies Based on Large Health Care Databases: Problems and Potential Solutions – a Primer for the Clinician." *Clin Epidemiol*. 9:185–193. doi:[10.2147/clep.S129879](#).
74. [△]Wallach JD, Serghiou S, Chu L, Egilman AC, Vasiliou V, Ross JS, Ioannidis JPA (2020). "Evaluation of Confounding in Epidemiologic Studies Assessing Alcohol Consumption on the Risk of Ischemic Heart Disease." *BMJ Med Res Methodol*. 20:64. doi:[10.1186/s12874-020-0914-6](#).
75. [△]Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA (2009). "High-Dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data." *Epidemiology*. 20:512–522. doi:[10.1097/EDE.0b013e3181a663cc](#).
76. [△]Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018). "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal*. 21:C1–C68. doi:[10.1111/ectj.12097](#).
77. [△]Meinshausen N, Bühlmann P (2010). "Stability Selection." *J R Stat Soc Series B Stat Methodol*. 72:417–473. doi:[10.1111/j.1467-9868.2010.00740.x](#).

78. [△]Schisterman EF, Cole SR, Platt RW (2009). "Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies." *Epidemiology*. **20**:488–495. doi:[10.1097/EDE.0b013e3181a819a1](https://doi.org/10.1097/EDE.0b013e3181a819a1).
79. [△]VanderWeele TJ (2019). "Principles of Confounder Selection." *Eur J Epidemiol*. **34**:211–219. doi:[10.1007/s10654-019-00494-6](https://doi.org/10.1007/s10654-019-00494-6).
80. [△]Greenland S, Pearl J, Robins JM (1999). "Causal Diagrams for Epidemiologic Research." *Epidemiology*. **10**:37–48.
81. [△]Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T (2006). "Variable Selection for Propensity Score Models." *Am J Epidemiol*. **163**:1149–1156. doi:[10.1093/aje/kwj149](https://doi.org/10.1093/aje/kwj149).

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.