Using Big Data from Internet to Improve Land Use Transition Effects on Eco-Environment

Tong Liu[1,2], Shijun Wang[3], Zhangxian Feng[3], Hongji Yang[4], Lin Zou[5]*

[1] College of Earth Sciences, Jilin University, China
[2] School of Architecture and Urban Planning, Jilin Jianzhu University, China
[3] School of Geographical Sciences, Northeast Normal University, China
[4] School of Computing and Mathematic Sciences, University of Leicester, UK
[5] School of Computer Science and Informatics, De Montfort University, UK
Liutong@jlju.edu.cn, wangsj@nenu.edu.cn, fengzx092@nenu.edu.cn, hongji.yang@leicester.ac.uk, lin.zou@dmu.ac.uk

Abstract

The intricate relationship between land use and ecological changes is a critical area of eco-environmental research. However, there is a notable lack of quantitative, analytical studies that explore this relationship in depth. This study aims to bridge this gap by applying a contrastive learning approach to a practical land use application, utilizing big data. We analyzed 2,197 publications from the Web of Science Core Collection spanning from 1998 to 2022. Our methodology correlates textual and visual data, enhancing the categorization and thematic analysis of existing literature. The contrastive learning technique uncovers nuanced relationships between textual descriptions and visual data, providing a comprehensive view of research trends. Key findings include an increasing volume of publications, indicating growing interest and potential in this field; predominant focus areas such as the relationship between climate change and land use, dynamic land use modelling, urbanization impacts, land use management, ecosystem services, and biodiversity. Emerging research hotspots identified through contrastive learning include the interactions between urban thermal environments and land use changes, land use effects on biological communities and soil organic matter, and the relationship between ecological conditions and land use in vulnerable regions. This study not only maps the current research landscape but also employs a contrastive learning approach to predict future research directions, thus enriching the methodological framework for analyzing land use applications with big data.

Keywords: Contrastive learning, Text clustering, Bibliometric, Big data

1 Introduction

Land use, a cornerstone of human economic and social progress, serves as a mirror reflecting the dynamic interplay between societal evolution and environmental stewardship. The initiation of the Land Use/Land Cover Change (LUCC) program in the 1990s by the International Geosphere-Biosphere Program (IGBP) and the International Human Dimensions Program (IHDP) marked a watershed moment, catalyzing global scholarly engagement with the multifaceted impacts of human activity on the planet's ecological and socio-economic fabric. This burgeoning field of study has progressively embraced the intricate connections between anthropogenic endeavors and environmental sustainability, championing the concept of land use transition as a lens through which to examine these complex dynamics.

Amid escalating global environmental concerns, the quest for sustainable development - harmonizing economic growth, social advancement, and environmental preservation—has never been more pressing. This imperative resonates strongly in resource-rich locales, where the dichotomy between development and conservation presents stark challenges, notably the encroachment of construction on ecological sanctuaries. Such imbalances underscore the urgency of adopting rigorous analytical frameworks to assess and mitigate the adverse effects of unchecked land-use transitions.

For example, in China, a nation at the epicenter of profound land use alterations driven by rapid industrialization and urbanization, the adoption of the land use transition framework has ignited significant scholarly interest. This is particularly evident in studies spanning the Yangtze River transect, illustrating the critical need for a methodological pivot towards quantitative, evidence-based analyses to dissect the nuances of land use change and its implications.

Bibliometrics provides a robust framework for scrutinizing the wealth of research on land use transitions. By tracking literature trends, academic influence, and collaboration networks, bibliometric analysis offers a predictive vista into the trajectory of this field. The integration of advanced contrastive learning techniques, such as Contrastive Language-Image Pretraining (CLIP), further enhances this endeavor. CLIP allows for the simultaneous analysis of textual and visual data, offering a richer, multimodal understanding of research trends and thematic evolutions.

*Corresponding Author: Lin Zou; E-mail: lin.zou@dmu.ac.uk
DOI: https://doi.org/10.70003/160792642025012601007

This contrastive learning approach enables the visual elucidation of complex data and trends, thereby democratizing the understanding of land use transition research. By applying these advanced analytical techniques, this study aims to forge a deeper understanding of the thematic and conceptual evolutions within the land use transition discourse. This integration not only contributes to the global dialogue on sustainable land use management strategies but also enriches the methodological arsenal available for analyzing land use transitions.

2 Related Work

The study of land use transition and its environmental effects has garnered considerable attention, particularly with the advent of rapid agricultural and urban expansion. Significant transformations in farmland usage within the sample underscore the profound policy implications these changes harbor [1]. Such a transformation is not confined to China alone; the impacts of urbanization on rural land use and livelihoods have been globally recognized, as evidenced by the case study of Hanoi, Vietnam [2].

Moreover, the importance of green infrastructure and biodiversity within urban contexts has been highlighted through studies on Mediterranean urban parks [3], emphasizing the significance of historical, urban, managerial, and soil characteristics in preserving ecological balance. The essential role of advanced bibliometric tools in tracking research trends in this domain has been addressed [4], offering a contemporary review of software tools that facilitate the analysis of scientific advancements within this field.

Further exploration into the ramifications of land use transitions on soil nitrogen dynamics [5], alongside studies on the interactive effects of climate and land use on ecosystem services [6], delineates the intricate interplay between land use changes, environmental consequences, and biodiversity. Collectively, these studies contribute to a comprehensive understanding of how land use management strategies can sustain the dynamic equilibrium of ecological environments, emphasizing the need for integrated approaches to urban and subsurface management amidst the challenges posed by ongoing urbanization processes [7].

2.1 Literature Selection for Bibliometrics

Bibliometrics, as a methodological approach in scientific research, relies heavily on the systematic quantification of literature to analyze patterns, trends, and gaps within a specific field of study. The foundation of any bibliometric analysis is the corpus of literature selected for examination. This selection process is critical because the quality, relevance, and scope of the literature directly influence the validity and comprehensiveness of the analysis. Good literature selection ensures that the study accurately reflects the state of research, identifies the most influential works and authors, and captures the evolution of themes and methodologies over time [8-9].

Properly selected literature ensures comprehensive coverage of the topic, including seminal works, recent advancements, and emerging trends. It helps in delineating the research landscape, facilitating a more accurate and holistic understanding of the field. Moreover, by employing systematic and transparent selection criteria, researchers can minimize selection bias, enhancing the reliability and objectivity of the bibliometric analysis [10].

Advancements in digital databases and search algorithms have significantly improved the process of literature selection for bibliometric analyses. These technological advancements facilitate more efficient and precise retrieval of relevant publications, enabling researchers to construct a more accurate and comprehensive dataset for analysis. Key advancements include sophisticated search algorithms, automated filtering and sorting, cross-database searching, and citation network analysis tools [11].

The selection of literature is foundational to the success of bibliometric analyses, ensuring that studies are comprehensive, objective, and reflective of the current state of research. The advancements in digital databases and analytical tools have significantly enhanced the efficiency and precision of literature selection, enabling more sophisticated and nuanced analyses of scientific literature.

2.2 Clustering Research

In the realm of machine learning and data mining, clustering algorithms are fundamental tools designed to group a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups. These algorithms are widely used across various fields, such as computer science, biology, marketing, and social science, to discover natural groupings within data.

There are several types of clustering algorithms, each with its own methodology and application areas. The choice of algorithm depends on the dataset's characteristics and the specific requirements of the task. Below, we will discuss some of the most prominent clustering algorithms.

K-means is one of the simplest and most commonly used clustering techniques. It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Hierarchical clustering seeks to build a hierarchy of clusters. It can be divided into two main types: agglomerative (bottom-up approach) and divisive (top-down approach). Density-Based Spatial Clustering of Applications with Noise (DBSCAN) groups together closely packed points by identifying core points, border points, and noise. Gaussian Mixture Models (GMM) is a probabilistic model that assumes all the data points are generated from a mixture of several Gaussian distributions with unknown parameters. Spectral clustering techniques make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions [12].

In this research, k-nn clustering performs best for its simplicity and efficiency. It requires specifying the number of clusters in advance, which might necessitate some experimentation or domain knowledge to determine.

2.3 Bibliometric Methods

In the study of land use transition effects on the eco-environment, we employ a multifaceted analytical approach. This approach encompasses keyword network analysis, cluster analysis, timeline analysis, and burst analysis. Each technique offers unique insights into the bibliometric landscape, revealing the thematic, chronological, and emergent properties of the research domain.

Keyword network analysis is pivotal in visualizing thematic structures within a dataset, illustrating the connections among key concepts [13]. This analysis elucidates the central themes and their interrelationships, providing a map of the research field's intellectual territory. Cluster analysis categories entities such as publications or keywords into groups based on their similarities, highlighting the research field's subdomains. This technique is essential for identifying cohesive areas of study within the broader research landscape [14]. Timeline analysis offers a longitudinal view of the research field, tracking the evolution of themes over time. It reveals the historical development of topics, identifying shifts in focus and emerging areas of interest [15]. Burst analysis detects sudden increases in the frequency of specific terms, signaling emerging trends or shifts in research emphasis. This method is crucial for identifying cutting-edge topics within the field [16]. Machine learning techniques can significantly enhance these bibliometric analyses, providing more sophisticated tools for pattern recognition, trend forecasting, and thematic clustering [17].

These analytical methods form a comprehensive toolkit for bibliometric studies, each contributing a unique lens through which the vast landscape of literature on land use transition and its eco-environmental effects can be examined.

2.4 Contrastive Learning

Contrastive learning has emerged as a powerful technique in unsupervised and self-supervised learning, primarily aimed at learning effective feature representations by comparing positive and negative pairs of data. This method, which has gained significant attention in the field of computer vision and natural language processing, aligns well with various applications including image classification, retrieval, and zero-shot learning.

The researchers introduced a simple yet effective framework for contrastive learning of visual representations, demonstrating that contrastive learning can outperform traditional supervised learning methods in many tasks by leveraging large-scale, unlabeled data. This approach utilizes a contrastive loss to bring similar data points closer in the embedding space while pushing dissimilar ones apart [18].

Similarly, Radford developed the CLIP (Contrastive Language-Image Pretraining) model, which leverages a vast amount of (image, text) pairs collected from the internet. CLIP aligns image and text embeddings in a shared space, enabling powerful zero-shot learning capabilities that allow the model to generalize across a wide range of visual concepts and tasks without the need for task-specific training [19].

Other researchers extended the contrastive learning paradigm with the introduction of Momentum Contrast (MoCo), a method that builds a dynamic dictionary with a queue and a moving-averaged encoder, addressing the memory and computational constraints of previous contrastive methods [20].

These advancements in contrastive learning provide a solid foundation for exploring its application to land use transition analysis, offering robust tools for identifying trends, hotspots, and future directions in land use changes.

3 Main Approach

3.1 Overview

Contrastive Language-Image Pretraining (CLIP) represents a significant advancement in the field of machine learning, enabling the simultaneous analysis of textual and visual data. This approach utilizes a large-scale dataset of images and their corresponding textual descriptions to train a model that can understand and relate these modalities. In our research, CLIP is employed to enhance the bibliometric analysis of land use transition literature by providing a more holistic view of research trends. By linking textual data with visual representations, CLIP facilitates the identification of nuanced relationships and thematic evolutions within the corpus of publications. This multimodal analysis not only deepens the understanding of land use transition dynamics but also aids in the prediction of future research directions, thereby enriching the methodological framework for examining the ecological implications of land use changes.

Figure 1 and Figure 2 below illustrate the process of Contrastive Language-Image Pretraining (CLIP).

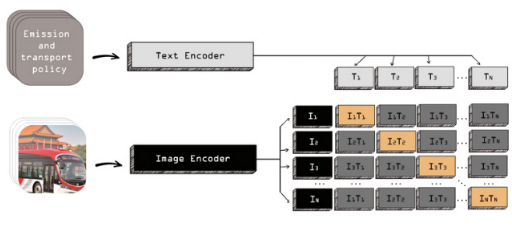

Figure 1. Contrastive learning

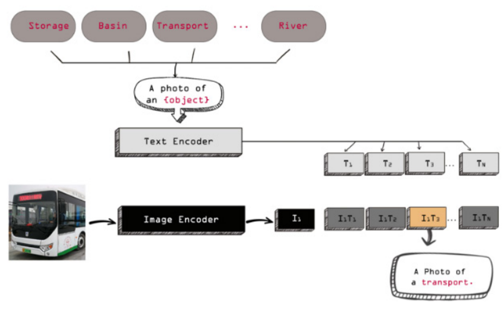

Figure 2. Prediction process

3.1.1 Contrastive Pre-training

Text Encoder: The text encoder processes textual descriptions (e.g., "Pepper the Aussie pup") and converts them into a corresponding set of text embeddings.

Image Encoder: The image encoder processes images (e.g., an image of a dog) and converts them into a corresponding set of image embeddings.

Contrastive Learning: The embeddings from the text and image encoders are then paired, and a contrastive loss is applied. This loss function ensures that the embeddings of corresponding images and text are close to each other in the embedding space, while embeddings of non-matching pairs are far apart. This alignment process allows the model to learn the relationship between textual descriptions and visual content.

3.1.2 Create Dataset Classifier from Label Text

Label Text Encoding: Text descriptions of various labels (e.g., plane, car, dog, bird) are processed by the text encoder to create text embeddings for each label.

Class Embeddings: These text embeddings represent the different categories that the model will classify images into.

3.1.3 Use for Zero-shot Prediction

Image Encoding for Prediction: A new image (e.g., an image of a dog) is processed by the image encoder to generate its image embedding.

Prediction: The generated image embedding is compared against the precomputed text embeddings of the labels. The model identifies the label whose text embedding is closest to the image embedding. In this example, the model predicts "a photo of a dog" as the label that best matches the new image.

Overall, the CLIP model leverages contrastive learning to align textual and visual information, enabling it to understand and classify images based on textual descriptions without needing explicit training on the specific classification task (zero-shot learning). This process allows for a flexible and powerful way to perform image classification using textual labels.

3.2 Data Collection Source

Deep clustering transcends traditional text analysis techniques by addressing the inherent limitations associated with the "curse of dimensionality" and the sparse, high-dimensional nature of textual data. This approach ensures a more accurate and insightful grouping of text documents, revealing the underlying thematic structures and facilitating a granular understanding of land use transition research.

Within the realm of deep clustering for text analysis, the k-nearest neighbors (k-NN) algorithm is instrumental in refining the clustering process. After the transformation of textual data into a lower-dimensional feature space through deep learning techniques, k-NN is utilized to explore the local neighborhoods of embedded texts. This step is crucial for several reasons.

Precision in Thematic Grouping: By identifying the 'k' most similar documents to a given text, k-NN aids in ensuring that the clusters formed are coherent and thematically consistent, enhancing the precision of the analysis.

Mitigation of Noise: The application of k-NN contributes to the reduction of noise within the clustered datasets, as it averages the features within local neighborhoods, thereby smoothing out anomalies and outliers.

Enrichment of Semi-supervised Learning: In in- stances where partial labels are available, k-NN can leverage this information to improve the clustering outcome, making the deep clustering process even more robust and informed.

In this research, given a dataset of document vectors transformed into a lower-dimensional feature space by a deep learning model, where each represents the deep-learned features of document pertaining to land use transition research, the tailored function of the k-NN algorithm proceeds as follows (1):

$$d(D, D_i) = \sqrt{\sum_{j=1}^{m}(D_j - D_{ij})^2} \quad \textbf{(1)}$$

Where m is the number of dimensions in the feature space d(D, D_i) calculates the Euclidean distance between the document vector D and another document vector D_i, which is critical for determining the proximity and thematic similarity between documents.

$$C = \text{mode}\{C_i | (D_i, C_i) \in k\text{-nearest of } D\} \quad \textbf{(2)}$$

C represents the cluster assignment for document \$D\$, determined by the majority cluster among its k nearest neighbors in (2) function. This step is crucial for ensuring that documents within the same cluster exhibit high thematic relevance to land use transition research.

This tailored approach underscores the methodological rigor applied in employing the k-NN algorithm to enhance deep text clustering, specifically aimed at uncovering thematic structures within the corpus of land use transition literature.

This process for this study was meticulously designed to harness the vast repository of knowledge contained within the Web of Science Core Collection. This esteemed database encompasses over 18,000 authoritative journals spanning a wide array of disciplines, including but not limited to natural sciences, engineering, biomedicine, social sciences, arts, and humanities. Central to its utility is a unique citation index, meticulously organized by cited authors, sources, and publication years, making it an indispensable tool for scholarly research worldwide.

Our data extraction was guided by precisely formulated search queries aimed at capturing the broad spectrum of research concerning land use transitions and their ecological impacts, with a specific focus on China. The search strategy employed the query by using the following equation (3):

$$Q = L \wedge (A \vee C \vee T) \wedge (E \vee Ec \vee I \vee Ic) \tag{3}$$

Where L is land use A is after C is change, T is transition, E is environment effect, Ec is ecology effect, I is environment impact, and Ic is ecology impact.

Q is the entire query structure; L initiates the search with a broad focus on land use; A.C, and T expand the query to include documents that discuss any form of alteration, change, or transition related to land use; E, Ec, I, and Ic ensure the retrieval of documents that cover the effects or impacts on the environment and ecology due to land use practices.

The inclusion of a fussy string, denoted by "*", was pivotal in broadening the search to encompass a wide range of terminologies pertinent to the study's keywords. On April 27, 2022, this comprehensive search strategy yielded an initial retrieval of 12,710 documents. Through a rigorous selection process, prioritizing articles that specifically address the context of China, a dataset comprising 2,197 documents was meticulously compiled for in-depth analysis.

This substantial dataset served as the foundation for our subsequent analytical endeavors. Applying a combination of keyword network analysis, cluster analysis, timeline analysis, and burst analysis, as discussed in the "Analysis Methods" section, we delved into the thematic structures, temporal evolutions, and emergent trends within the corpus. Each technique, bolstered by the integration of machine learning algorithms, offered unique insights into the complex landscape of land use and environmental research. This analytical synergy not only facilitated a comprehensive understanding of the existing literature but also highlighted the dynamic interplay between land use practices and ecological outcomes.

3.3 InfoNCE based Loss Function

The contrastive loss function is pivotal in our research and the CLIP model, as it fundamentally drives the alignment between textual and visual data. Specifically, the contrastive loss function used in CLIP is based on the InfoNCE (Information Noise-Contrastive Estimation) concept. This function operates by maximizing the similarity between positive pairs (matching image and text) while minimizing the similarity between negative pairs (non-matching image and text). This dual objective ensures that the model learns to distinguish relevant features that link textual descriptions to their corresponding images accurately. In our study on land use transition, employing this loss function allows for a nuanced understanding of how textual research outputs relate to visual representations of land use changes. This methodology enhances the robustness of our bibliometric analysis, providing deeper insights into the thematic and conceptual connections within the literature, thus enriching the analytical framework for assessing ecological implications of land use transitions.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(\mathrm{sim}(\mathbf{I}_i,\mathbf{T}_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\mathbf{I}_i,\mathbf{T}_j)/\tau)} + \log\frac{\exp(\mathrm{sim}(\mathbf{T}_i,\mathbf{I}_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\mathbf{T}_i,\mathbf{I}_j)/\tau)}\right] \tag{4}$$

In Function (4) where N is the number of (image,text) pairs, the image and text embeddings for the i-th pair, respectively, sim denotes the cosine similarity between two embeddings. In our approach, the contrastive loss is calculated using the above formula, which encourages matched pairs to have high similarity and unmatched pairs to have low similarity.

3.4 Training Procedure and Analysis of Publication Trends

Next step, our research approaches a structured algorithm (1) for the CLIP training procedure, covering data preprocessing, forward pass, similarity calculation, loss calculation, and the optimisation process. It also includes the mathematical representation of the contrastive loss function.

Utilising defined parameters, we identified 2,197 publications relevant to land use transitions, spanning from January 1, 1998, to April 27, 2022. A significant majority, 97.31% (2,138 entries), were categorised as research articles, showcasing the scholarly community's emphasis on empirical and theoretical studies within this domain. The remaining 2.69 comprised early access publications, proceeding reports, and book chapters, highlighting a diverse yet focused array of scholarly outputs [16, 21]. Types of documents published of the research about LUT effects on eco-environmental is showed in Figure 3.

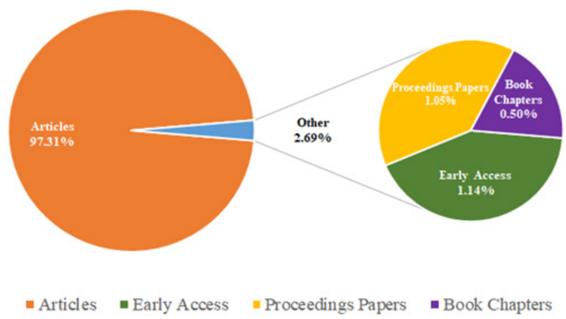

Figure 3. Types of documents

An annual publication trend analysis revealed a general upward trajectory in the volume of literature, with a noteworthy spike of 418 articles in 2021, indicating an escalating interest in this field [22]. This trend facilitates the division of research activity into four distinct phases, reflecting shifts in focus and methodology over time, from the initial introduction of land use transformation concepts to the current period of rapid growth and heightened scholarly attention [23-24].

Algorithm 1. Training procedure for CLIP

Input: Dataset of N (image, text) pairs $\{(\mathbf{I}_i, \mathbf{T}_i)\}_{i=1}^{N}$
Parameters: Temperature parameter $\mathcal{T}$, learning rate η, number of epochs E
Initialise image encoder $f_{\text{img}}(\cdot)$ and text encoder $f_{\text{text}}(\cdot)$
for epoch = 1 to E **do**
  for each batch of (image, text) pairs **do**
    Data Preprocessing:
    Normalise and resise images
    Tokenise and convert text to numerical representations
    Forward Pass:
    Compute image embeddings $\mathbf{z}_i = f_{\text{img}}(\mathbf{I}_i)$ for $i = 1, ..., N$
    Compute text embeddings $\mathbf{t}_i = f_{\text{text}}(\mathbf{T}_i)$ for $i = 1, ..., N$
    Similarity Calculation:
    Compute cosine similarity matrix $S_{ij} = \text{sim}(\mathbf{z}_i, \mathbf{t}_j)$
    Loss Calculation:
    Calculate contrastive loss
    Backward Pass and Optimisation:
    Compute gradients of $\mathcal{L}$ with respect to the model parameters
    Update model parameeters using optimiser (e.g. Adam) with learning rate η
  end for
end for
Output: Trained image encoder $f_{\text{img}}(\cdot)$ and text encoder $f_{\text{text}}(\cdot)$

Figure 4 presents the number of published documents of the research about LUT effects on eco-environmental. Moreover, the dual-map overlay technique, as refined by, was employed to elucidate the disciplinary reach and citation dynamics within this corpus, revealing a strong concentration in ecology, earth, and marine sciences, among others. This multidisciplinary engagement underscores the complex and interrelated nature of land use transition studies, suggesting an evolving landscape of research that spans across various scientific domains [25].

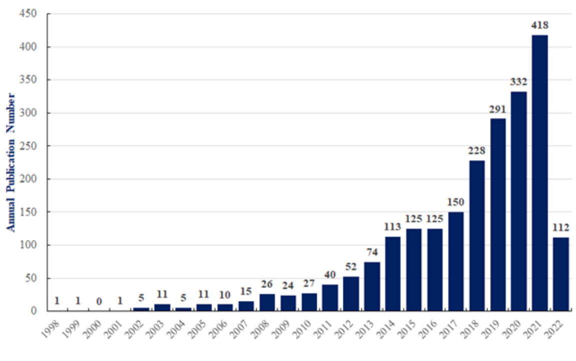

Figure 4. The number of published documents

Furthermore, an examination of publication and citation frequencies within this dataset revealed key journals that significantly influence the field in Table 1, underscoring the importance of interdisciplinary collaboration in advancing our understanding of land use transitions and their environmental impacts [26].

Analysis of institutional contributions highlighted the Chinese Academy of Sciences' central role, yet revealed relatively isolated research clusters, suggesting an opportunity for enhanced cooperation among leading institutions [27-28].

This narrative is supported by author network analyses conducted via CiteSpace, which pinpointed pivotal contributors and collaborative patterns within the field, offering insights into potential research hotspots and future directions.

Table 1. The top 10 journals with eco-environmental effects of LUT research

Journal	Year	Cent	Frequency
1 SCI TOTAL ENVIRON	1999	0.41	986
2 SCIENCE	2001	0.2	968
3 NATURE	1999	0.4	902
4 P NATL ACAD SCI USA	2008	0	709
5 ECOL INDIC	2014	0.1	691
6 J ENVIRON MANAGE	2006	0.05	665
7 ARG ECOSYS ENVIRON	2002	0.23	642
8 LAND USE POLICY	2012	0.04	610
9 LANDSCAPE URBAN PLAN	2006	0.05	589
10 GLOBAL CHANGE BIOL	1999	0.04	561

3.5 Data Representation and Tools

CiteSpace is an indispensable software tool designed for the visual analysis of bibliometric data and the identification of emergent trends within specific fields of research. Its compatibility with multiple bibliographic databases, including the Web of Science (WoS), Scopus, and the Chinese Social Science Citation Index (CSSCI), empowers researchers to undertake comprehensive analyses of new trends and knowledge domains. This software's distinctive capabilities extend to geographical visualization, allowing for an in-depth exploration of geospatial correlations and facilitating a nuanced understanding of spatially distributed academic discourse; our study embarked on a bibliometric exploration to unravel the intricacies of land use transition and its consequential effects on the environment. The analytical process encompassed the examination of various scholarly outputs, spanning the type and volume of publications, contributing authors and institutions, influential journals, prevalent keywords, and citation dynamics in Figure 5.

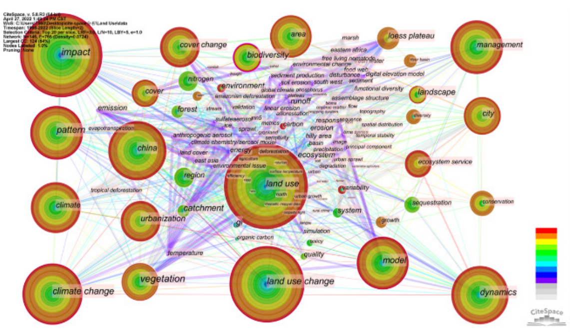

Figure 5. Keywords network

The analysis of keywords serves as a gateway to apprehending the overarching themes, methodologies, and objectives permeating the literature. By conducting a detailed co-occurrence and cluster analysis of keywords, we aim to dissect the core research foci and methodological orientations that define this field. CiteSpace's capability to automatically aggregate and synthesize keyword data is pivotal, with the frequency of keyword occurrence offering insights into the thematic priorities and conceptual frameworks guiding land use transition research. Notably, “land use” and “impact” emerge as predominant keywords, underscoring the centrality of these concepts in scholarly investigations into the ecological and environmental ramifications of land use dynamics. This revelation aligns with the broader academic endeavor to elucidate the multifaceted interactions between land use practices and their environmental outcomes [29-30].

Furthermore, the recurring emphasis on “climate change” within the keyword spectrum highlights the critical intersection between land use transitions and global climatic alterations. The substantial body of research dedicated to this nexus reflects a concerted effort to unravel the contributions of land use changes to greenhouse gas emissions and climate variability, underscoring the urgency of integrating land use planning with climate change mitigation strategies [5, 31]. Similarly, the focus on “dynamics” and “patterns” within the keyword analysis reveals a scholarly inclination towards quantifying land use changes and understanding their implications for food security, biodiversity conservation, and climate resilience. This analytical perspective is instrumental in crafting informed land management policies and sustainable development frameworks that address the complex challenges posed by land use transitions [32-33].

In summary, our bibliometric analysis, facilitated by CiteSpace, offers a panoramic view of the evolving landscape of research on land use transition and its environmental effects. Through the lens of keyword co-occurrence and cluster analysis, we delineate the primary thematic contours and methodological approaches that characterize this field of study. The insights garnered from this analysis not only illuminate the current state of research but also chart potential trajectories for future investigations, thereby contributing to the formulation of holistic and sustainable land use and environmental management strategies.

4 Result and Discussion

4.1 Keywords Cluster

Cluster analysis of keywords is a technique for organizing a vast array of keywords into distinct clusters based on their interrelatedness, thereby unveiling the predominant research directions within a field. This method was applied and depicted using CiteSpace, illustrating the distribution and relational dynamics of keywords across various clusters in Figure 6. The numerical labelling of clusters serves as an inverse indicator of keyword density within each cluster, with more compact numbers signaling a higher concentration of keywords. The efficacy of the clustering process is quantitatively assessed through Modularity Q and Silhouette S metrics, where higher Q values signify enhanced clustering quality. A threshold of Q greater than 0.3 and S exceeding 0.5 denotes a clustering outcome of commendable accuracy. The analysis, as represented, indicates a robust clustering structure, affirming the coherence and analytical robustness of the resultant clusters.

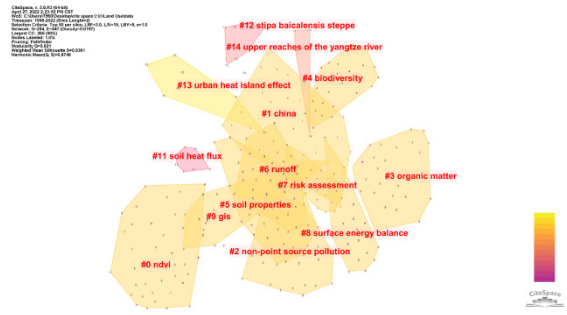

Figure 6. Keywords cluster

Within the context of land use transition and its environmental implications in China, 14 distinct clusters were identified. Among these, the clusters characterized by the Normalized Difference Vegetation Index (NDVI), non-point source pollution, and organic matter emerged as the most keyword-dense, highlighting their prominence in the research landscape. The central clustering of soil properties, runoff, and risk assessment keywords underscores a significant thematic overlap, suggesting these areas are intrinsically linked within the realm of current research inquiries. Notably, risk assessment emerges as a pivotal theme, evidencing its integral role in the environmental impact assessment of land use practices.

The analysis further revealed the strategic positioning of Geographic Information System (GIS) technology as an essential tool for managing natural resources and monitoring land use alterations. The utility of GIS and satellite imagery in tracing the long-term evolution of land use and its ecological consequences exemplifies the technology’s critical function in environmental restoration endeavors. Moreover, the peripheral positioning of keywords such as NDVI, organic matter, biodiversity, soil heat flux, and the urban heat island effect indicates their relatively specialized focus within the broader research spectrum. The identification of Stipa steppe and the upper reaches of the Yangtse River as emerging areas of interest points towards potential shifts in geographical research.

This nuanced keyword cluster analysis, devoid of specific bibliometric citations, affords a comprehensive overview of the thematic structures and evolving trends within the study of land use transition and environmental effects, highlighting the utility of advanced analytical tools in capturing the complexity of research dynamics.

4.2 Timeline Analysis

Hot topics within research fields evolve over time, necessitating an examination of shifts in research focal points to gain a more nuanced understanding of the leading edges and developmental trajectories within a domain. In Figure 7, the deployment of CiteSpace5.8.R3’s functionality to segregate clustering outcomes by year facilitated the creation of a temporal axis diagram, mapping the chronological clustering of keywords. This approach underscores the changing landscape of research emphases, where the size of each node in the diagram correlates with the frequency of keyword appearances, illustrating the prominence of specific topics at various junctures.

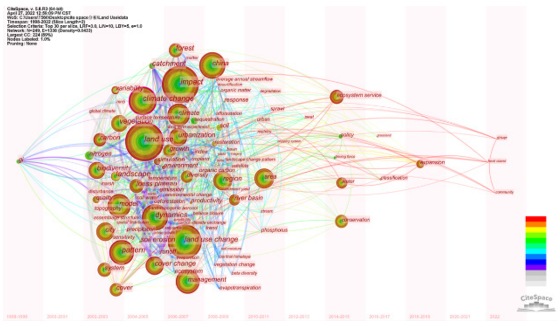

Figure 7. Timeline map

The inception of Green Infrastructure (GI) as a keyword, dominating the discourse from 1998 to 2002, signifies the enduring relevance of this concept. Originating in the United States towards the twentieth century's close, GI employs a nature-based schema utilizing vegetation, soil, and permeable surfaces to enhance rainwater management and establish interconnected green spaces, thereby mitigating urban heat island effects and promoting environmental sustainability. This philosophy underscores the integration of ecological considerations into urban planning, epitomizing a shift towards sustainability in land use management.

Subsequent years witnessed a diversification of research hotspots, with a notable increase in high-frequency keywords from 2002 to 2011. This period was characterized by a focus on developing models to track the dynamic interplay between land use patterns and climatic variables, examining the ramifications of land use alterations, and exploring the ecological impacts induced by rapid urbanization. The emergence of themes related to ecosystem services, conservation efforts, water resource management, policy formulation, and driving forces between 2014 and 2015 highlights growing scholarly attention towards the valuation of ecosystem services and the influence of land use on hydrological environments.

Recent trends indicate a pivot towards examining the socio-ecological drivers of land use and ecological changes, underscored by urbanization and urban sprawl, and their implications on biodiversity and the soil microbiome. This shift reflects an increasing recognition of human agency in shaping land use dynamics and the critical need for integrated approaches to manage the intersection of urban development and environmental stewardship.

4.3 Burst Detection

Keywords that emerge frequently within a specified timeframe, albeit with a lower occurrence rate, are identified as burst words. These burst words signify the rapid emergence of specific research themes, providing insights into the evolving trends within the field of study. The application of CiteSpace's burst item detection tool facilitated a visual examination of keywords spanning from 1998 to 2022, enabling the identification of the top 30 keywords demonstrating the most significant burst activity over periods extending beyond two years.

The visual representation outlined in the Figure 8 highlights temporal intervals with each segment corresponding to a year, and the periods of burst word emergence are depicted in red in Figure 6. Notably, Green Infrastructure (GI) emerges as the keyword with both the earliest appearance and the longest duration of interest, sustaining relevance for 13 years. A pattern of elongated burst durations for keywords appearing before 2014 transitions to shorter durations from 2015 to 2018, indicating a period of rapid shifts in research focuses within the domain of land use transition and its environmental effects. From 2019 onwards, an elongation in the burst word cycles suggests a consolidation and stabilization of research directions in recent years.

Distinct variances in emergent keywords across different periods underscore the dynamic shifts in research emphases over time. Initial studies, spanning from 1998 to 2010, predominantly concentrated on the environmental impacts of land use in geographically diverse and ecologically sensitive areas such as hilly regions, cultivated lands, basins, and desertification-prone sones. Subsequent discussions from 2012 onwards expanded to encompass the effects of land use on soil organic matter and carbon sequestration, alongside examinations of human-induced land use transformations, including urban sprawl and the influence of policy and built environment dynamics.

Persisting burst keywords such as NDVI, ecosystem service, expansion, river basin, Loess Plateau, and temperature reflect sustained research interests. These focal areas include the application of NDVI for vegetation coverage assessment, evaluation of ecological environment changes, the ramifications of urban and population expansion on land use and ecological conditions, and the exploration of land use impacts on ecosystem services. Additionally, studies delve into the ecological and environmental dynamics of specific geographical locales such as river basins and the Loess Plateau, alongside investigations into the interplay between urban thermal environments, surface temperature variations, and land use alterations.

4.4 Discussion

Countries must modernize the harmonious coexistence of man and nature and promote sustainable economic and social development, requiring coordinated development of land use and the ecological environment. Therefore, theoretical innovation and technological updates on land use transformation and ecological and environmental effects are urgently needed. As shown above, there has been a lot of work in this field and significant progress has been made, but there are still some aspects that need further improvement and development.

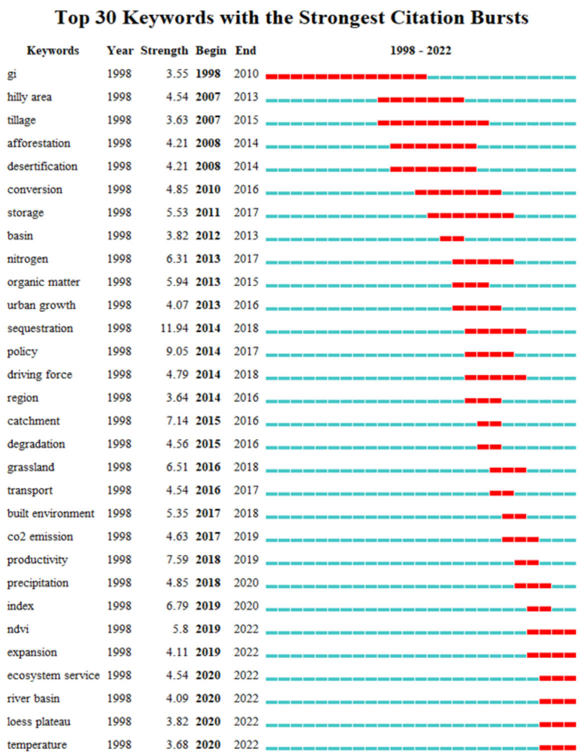

Figure 8. The top 30 results

The relationship between land use transformation and ecological environmental effects is relatively complex, involving disciplines such as geography, environmental science, sociology, and remote sensing science. Different environmental factors have different effects on different components, such as atmosphere, soil, and hydrology. A more comprehensive understanding of ecological and environmental effects requires the acquisition of accurate data and interdisciplinary work.

There are many driving factors for land use transformation and ecological environment changes, and the measurement indicators of dominant morphological changes are easier to quantify. Future work should focus on the development of improved analysis methods and technologies for the observation and monitoring of less obvious morphological changes to facilitate further quantitative analysis.

The urbanization process often has negative ecological effects, and the ecological environment may adversely affect land use. By better understanding changes in the ecological environment, land use strategies could be adjusted for more efficient land use transformation during urban-rural integration. Improved models are required to determine the best land resource management policies. Models should consider changes in the ecological environment for the timely adjustment of land use strategies, allowing for positive feedback from the ecological environment system, for effective land use transformation in the process of urban-rural integration.

5 Result and Discussion

The theory of land use transition, since its introduction to the Chinese academic discourse in 2001, has spurred a comprehensive and increasingly interdisciplinary exploration, particularly highlighted in the prolific period of 2002 to 2011. The subsequent phase, post-2012, marks a maturation in the research trajectory, with scholarly efforts consolidating around deep dives into critical issues and leveraging advanced technological methodologies.

Our investigation, harnessing the power of bibliometric analysis complemented by advanced contrastive learning techniques, has unearthed pivotal insights into the evolving landscape of land use transformation research in China and its profound ecological implications. The analysis has brought to the fore significant themes such as the ecological impacts of varied land use practices on soil organic matter, ground temperature, and surface runoff, thereby underscoring the imperative of judicious land use management for ecological conservation. The evaluation of ecosystem services, NDVI, and landscape patterns has been identified as crucial for the ecological assessment of land use, spotlighting the criticality of such assessments in sustainable land management.

Moreover, our research highlights focused inquiries into urban surface runoff, soil quality, and non-point source pollution, aiming at a quantitative understanding of land use practices' environmental repercussions. This body of work lays a theoretical groundwork for the strategic allocation of land resources to foster ecological balance. Issues such as urban heat islands, the interplay between urban rainstorms, floods, and land use changes further delineate the environmental challenges posed by urbanization and urban expansion.

The spotlight on ecologically fragile areas undergoing rapid land use changes, including the Loess Plateau and key river basins, calls for the development of sustainable land use planning and management strategies. Such strategies are pivotal in maintaining ecological integrity amidst evolving land use paradigms.

Employing bibliometric methods and advanced contrastive learning techniques, this study has mapped the intellectual landscape of land use transformation research, identifying influential scholars, institutions, and journals, and delineating research hotspots and frontiers. This comprehensive analysis not only reveals the dynamic evolution of the research field but also forecasts emergent trends, offering a scaffold for future scientific inquiries into land use transition and its environmental ramifications.

As the volume of publications continues to ascend, it underscores the burgeoning significance of this research domain within the broader contours of land system and environmental science. The insights garnered through our methodological framework enrich the knowledge base, enabling researchers to navigate the complex terrain of land use transition research more effectively and to pinpoint areas ripe for future exploration.

In conclusion, this study not only charts the historical and thematic developments in land use transition research but also showcases the utility of contrastive learning techniques in identifying and analyzing research trends. It sets the stage for a nuanced, informed, and forward-looking engagement with land use and environmental sustainability challenges, heralding a new era of integrated, technology-driven environmental research.

Acknowledgements

This research was partially funded by the National Natural Science Foundation of China (Grant No. 42071219), Science and Technology Plan Project of Jilin Province (Grant No. 20210203146SF and No. 202403044142SF), Scientific research project of the education department of Jilin Province (Grant No. JJKH20210284KJ), Jilin Education Science Planning Project (Grant No. ZD24046).

References

[1] L. Ma, H. Long, S. Tu, Y. Zhang, Y. Zheng, Farmland Transition in China and its Policy Implications, *Land Use Policy*, Vol. 92, Article No. 104470, March, 2020.

[2] Q. Nguyen, D. Kim, Reconsidering Rural Land Use and Livelihood Transition under the Pressure of Urbanization in Vietnam: A Case Study of Hanoi, *Land Use Policy*, Vol. 99, Article No. 104896, December, 2020.

[3] V. Milano, J. Cortet, D. Baldantoni, A. Bellino, F. Dubs, J. Nahmani, S. Strumia, G. Maisto, Collembolan Biodiversity in Mediterranean Urban Parks: Impact of History, Urbanization, Management and Soil Characteristics, *Applied Soil Ecology*, Vol. 119, pp. 428–437, October, 2017.

[4] J. A. Moral-Muñoz, E. Herrera-Viedma, A. Santisteban-Espejo, M. J. Cobo, Software Tools for Conducting Bibliometric Analysis in Science: An Up-to-date Review, *El Profesional de la Información*, Vol. 29, No. 1, Article No. e290103, January, 2020.

[5] M. Ojoyi, O. Mutanga, J. Odindi, J. Kahinda, E. Abdel-Rahman, Implications of Land Use Transitions on Soil Nitrogen in Dynamic Landscapes in Tanzania, *Land Use Policy*, Vol. 64, pp. 95–100, May, 2017.

[6] M. Peters, A. Hemp, T. E. Appelhans, J. N. Becker, C. Behler, A. Classen, F. Detsch, A. Ensslin, S. W. Ferger, S. B. Frederiksen, F. Gebert, F. Gerschlauer, A. Gütlein, M. Helbig-Bonitz, C. Hemp, W. J. Kindeketa, A. Kühnel, A. V. Mayr, E. Mwangomo, C. Ngereza, H. K. Njovu, I. Otte, H. Pabst, M. Renner, J. Röder, G. Rutten, D. S. Costa, N. Sierra-Cornejo, M. G. R. Vollstädt, H. I. Dulle, C. D. Eardley, K. M. Howell, A. Keller, R. S. Peters, A. Ssymank, V. Kakengi, J. Zhang, C. Bogner, K. Böhning-Gaese, R. Brandl, D. Hertel, B. Huwe, R. Kiese, M. Kleyer, Y. Kuzyakov, T. Nauss, M. Schleuning, M. Tschapka, M. Fischer, I. Steffan-Dewenter, Climate-land-use Interactions Shape Tropical Mountain Biodiversity and Ecosystem Functions, *Nature*, Vol. 568, No. 7750, pp. 88–92, April, 2019.

[7] L. Maring, M. Blauw, Asset Management to Support Urban Land and Subsurface Management, *Science of The Total Environment*, Vol. 615, pp. 390–397, February, 2018.

[8] M. Aria, C. Cuccurullo, Bibliometrix: An R-tool for Comprehensive Science Mapping Analysis, *Journal of Informetrics*, Vol. 11, No. 4, pp. 959–975, November, 2017.

[9] J. Mingers, L. Leydesdorff, A Review of Theory and Practice in Scientometrics, *European Journal of Operational Research*, Vol. 246, No. 1, pp. 1–19, October, 2015.

[10] L. Bornmann, R. Mutz, Growth Rates of Modern Science: A Bibliometric Analysis based on the Number of Publications and Cited References, *Journal of the Association for Information Science and Technology*, Vol. 66, No. 11, pp. 2215–2222, November, 2015.

[11] N. J. van Eck, L. Waltman, Visualizing Bibliometric Networks, in: Y. Ding, R. Rousseau, D. Wolfram (Eds.), *Measuring Scholarly Impact: Methods and Practice*, Springer, 2014, pp. 285–320.

[12] J. Cai, J. Hao, H. Yang, X. Zhao, Y. Yang, A Review on Semi-supervised Clustering, *Information Sciences*, Vol. 632, pp. 164–200, June, 2023.

[13] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, March, 2003.

[14] A. K. Jain, Data Clustering: 50 Years beyond K-means, *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651–666, June, 2010.

[15] R. J. Hyndman, G. Athanasopoulos, *Forecasting: Principles and Practice*, OTexts, 2018.

[16] C. Chen, *The CiteSpace Manual*, December, 2014. http://cluster.ischool.drexel.edu/~cchen/citespace/CiteSpaceManual.pdf

[17] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science and Business Media, 2009.

[18] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, *International Conference on Machine Learning*, Virtual Event, 2020, pp. 1597–1607.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models from Natural Language Supervision, in *International Conference on Machine Learning*, Virtual, 2021, pp. 8748–8763.

[20] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum Contrast for Unsupervised Visual Representation Learning, *Proceedings of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*, Seattle, WA, USA, 2020, pp. 9729–9738.

[21] C. Chen, Citespace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature, *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 3, pp. 359–377, February, 2006.

[22] C. Chen, Y. Chen, M. Horowitz, H. Hou, Z. Liu, D. Pellegrino, Towards An Explanatory and Computational Theory of Scientific Discovery, *Journal of Informetrics*, Vol. 3, No. 3, pp. 191–209, July, 2009.

[23] X. Chen, L. Yu, Z. Du, Y. Xu, J. Zhao, H. Zhao, G. Zhang, D. Peng, P. Gong, Distribution of Ecological Restoration Projects Associated with Land Use and Land Cover Change in China and Their Ecological Impacts, *Science of The Total Environment*, Vol. 825, Article No. 153938, June, 2022.

[24] L. Cuo, Y. Zhang, Y. Gao, Z. Hao, L. Cairang, The Impacts of Climate Change and Land Cover/use Transition on the Hydrology in the Upper Yellow River Basin, China, *Journal of Hydrology*, Vol. 502, pp. 37–52, October, 2013.

[25] G. Jin, X. Deng, X. Chu, Z. Li, Y. Wang, Optimization of Land-use Management for Ecosystem Service Improvement: A Review, *Physics and Chemistry of the Earth*, Vol. 101, pp. 70–77, October, 2017.

[26] R. Leimu, J. Koricheva, What Determines the Citation Frequency of Ecological Papers? *Trends in Ecology & Evolution*, Vol. 20, No. 1, pp. 28–32, January, 2005.

[27] L. Feng, S. Wang, H. Tang, X. Yu, S. Li, J. Yuan, G. Huang,

Knowledge Map of Rural Ecological Civilization Research based on CiteSpace V Visualization Analysis, *Journal of Ecology and Rural Environment*, Vol. 37, No. 2, pp. 137–144, February, 2021.
[28] X. Liu, X. Liang, X. Li, X. Xu, J. Ou, Y. Chen, S. Li, S. Wang, F. Pei, A Future Land Use Simulation Model (FLUS) for Simulating Multiple Land Use Scenarios by Coupling Human and Natural Effects, *Landscape and Urban Planning*, Vol. 168, pp. 94-116, December, 2017.
[29] V. Brovkin, L. Boysen, V. K. Arora, J. P. Boisier, P. Cadule, L. Chini, M. Claussen, P. Friedlingstein, V. Gayler, B. J. J. M. van den Hurk, G. C. Hurtt, C. D. Jones, E. Kato, N. de Noblet-Ducoudre, F. Pacifico, J. Pongratz, M. Weiss, Effect of Anthropogenic Land-use and Land-cover Changes on Climate and Land Carbon Storage in CMIP5 Projections for the Twenty-first Century, *Journal of Climate*, Vol. 26, No. 18, pp. 6859–6881, September, 2013.
[30] M. Howells, S. Hermann, M. Welsch, M. Bazilian, R. Segerström, T. Alfstad, D. Gielen, H. Rogner, G. Fischer, H. van Velthuizen, D. Wiberg, C. Young, R. A. Roehrl, A. Mueller, P. Steduto, I. Ramma, Integrated Analysis of Climate Change, Land-use, Energy and Water Strategies, *Nature Climate Change*, Vol. 3, pp. 621–626, July, 2013.
[31] T. Newbold, Future Effects of Climate and Land-use Change on Terrestrial Vertebrate Community Diversity under Different Scenarios, *Proceedings of the Royal Society B: Biological Sciences*, Vol. 285, No. 1881, Article No. 20180792, June, 2018.
[32] Y. Liu, H. Long, Land Use Transitions and Their Dynamic Mechanism: The Case of the Huang-Huai-Hai Plain, *Journal of Geographical Sciences*, Vol. 26, No. 5, pp. 515–530, May, 2016.
[33] K. Winkler, R. Fuchs, M. Rounsevell, M. Herold, Global Land Use Changes Are Four Times Greater than Previously Estimated, *Nature Communications*, Vol. 12, Article No. 2501, May, 2021.

Biographies

Tong Liu is a Ph.D in land resource management at College of Earth Sciences , Jilin University. Her main scientific research in the fields of land ecological security, land use change.

Shijun Wang is a professor, Ph.D supervisor at School of Geographical Sciences, Northeast Normal University. His main scientific research in the fields of urban geography, economic geography, and urban and rural planning.

Zhangxian Feng is an associate professor, Ph.D supervisor at School of Geographical Sciences , Northeast Normal University. His main scientific research in the fields of urban geography, shrinking cities, urban networks, and revitalization of Northeast China.

Hongji Yang is currently a Professor with the University of Leicester. His research interests include creative computing software engineering and Internet computing.

Lin Zou serves as a Lecturer at De Montfort University. He earned his PhD from the University of Leicester. His research primarily focuses on pragmatic data analysis within multimodal environments.