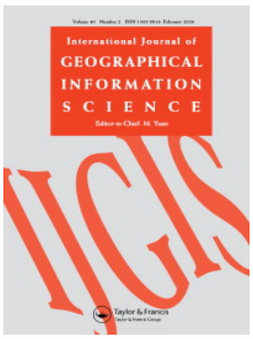

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

Second-dimension outliers for spatial prediction

Kai Ren, Yongze Song & Qiang Yu

To cite this article: Kai Ren, Yongze Song & Qiang Yu (24 Nov 2025): Second-dimension outliers for spatial prediction, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2025.2580414

To link to this article: https://doi.org/10.1080/13658816.2025.2580414

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2025.2580414

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Second-dimension outliers for spatial prediction

Kai Ren[a,b,c,d], Yongze Song[d] and Qiang Yu[b,c]

[a]State Key Laboratory of Climate System Prediction and Risk Management, Nanjing Normal University, Nanjing, China; [b]College of Natural Resources and Environment, Northwest A&F University, Yangling, Shaanxi, China; [c]State Key Laboratory of Soil and Water Conservation and Desertification Control, Northwest A&F University, Yangling, Shaanxi, China; [d]School of Design and the Built Environment, Curtin University, Perth, Australia

ABSTRACT

Spatial prediction aims to accurately estimate attributes at unsampled locations based on spatial dependencies, patterns, variability, and covariates, providing knowledge of complex spatial systems and supporting diverse applications. However, existing methods for spatial prediction ignore geographic and environmental characteristics outside sample locations, particularly spatial outliers, significantly impacting prediction accuracy. This study introduces the concept of second-dimension outliers (SDO) and SDO models that incorporate local outlier information at unsampled locations to enhance prediction accuracy. SDO models generate SDO variables that capture samples' external geographic and environmental characteristics in the spatial prediction. This study develops SDO-based machine learning to predict wheat production in Australia, using cross-validation to evaluate prediction accuracy. Results demonstrate that SDO-based support vector machines (SVM) improve spatial prediction accuracy, with the R^2 increasing from 0.555 to 0.671 compared to aspatial SVM, particularly for extreme values. The developed local outlier strength index that examines the strength of SDO ensures more accurate and smooth spatial predictions. The SDO concept provides more in-depth explanatory information from an innovative spatial perspective and a detailed understanding of local outliers for spatial prediction, making it a robust and effective tool for spatial statistical inference and geographic computation across various fields.

ARTICLE HISTORY
Received 14 November 2024
Accepted 22 October 2025

KEYWORDS
Spatial outliers; spatial prediction; second-dimension spatial association; agricultural production forecasting

1. Introduction

Spatial prediction aims at accurately estimating unknown values across a geographical area based on the spatial characteristics of attributes, such as spatial dependency, patterns, variability, and covariates. Spatial prediction is critically important across a range of fields, including Earth science, urban informatics, geosocial media analytics, agricultural management,

CONTACT Yongze Song yongze.song@curtin.edu.au; Qiang Yu yuq@nwafu.edu.cn.

meteorological disaster forecast, and public health (Sen 2016, Lu *et al.* 2017, Jiang 2019, Jia *et al.* 2023, Din and Yamamoto 2024, Zhang *et al.* 2024a). The rapid advancement of Earth observations and multi-source sensing technologies has resulted in the accumulation of extensive geospatial data, demonstrating the increased need for effective and efficient prediction methods (Mishra *et al.* 2017, Wadoux *et al.* 2019, Tian *et al.* 2022, Song *et al.* 2025). By employing dynamic spatial feature extraction and real-time spatial monitoring, these methods can fully explore large datasets to deliver accurate predictions and valuable spatial information (Vicente-Serrano *et al.* 2023, Yin *et al.* 2023). Effective predictive techniques enhance the understanding of spatial phenomena, facilitating informed decision-making and improved applications in diverse fields (Ulloa-Espindola and Perez-Albert 2022, Vicente-Serrano *et al.* 2023).

Methods for spatial prediction can be classified into the following categories, each with its unique strengths. The first category, aspatial models, includes traditional statistical methods and machine learning algorithms such as linear regression and random forest (Taheri Shahraiyni and Sodoudi 2016, Georganos and Kalogirou 2022). While these methods are effective for general prediction tasks, they often ignore spatial relationships, limiting their effectiveness in spatial contexts. The second category, spatial dependence models, includes techniques such as kriging approaches, spatial Bayesian hierarchical models, and geo-additive models (Strandberg *et al.* 2019, Wu *et al.* 2024). These models account for spatial variability by integrating different geographic characteristics and incorporating spatial autocorrelation to capture local variations. The third category, spatial heterogeneity models, features approaches that include geographically weighted regression (GWR) and its improved models (Harris *et al.* 2010, Tan *et al.* 2017). GWR specifically focuses on spatial variability and heterogeneity by examining how relationships between variables change across local locations. The fourth category, second-dimension spatial association (SDA) models, enhances spatial prediction by integrating spatial information from beyond sample locations, extracting second-dimension variables from explanatory data outside observed points to improve accuracy and capture deeper geographical patterns (Song 2022). Lastly, geographical similarity models use spatial configurations to predict variables based on their similarity to other locations, providing a flexible approach to managing diverse geographical contexts (Zhu *et al.* 2018, Song 2023). Each category demonstrates various advantages and addresses specific challenges in spatial prediction, contributing to a more in-depth understanding of spatial data.

Spatial outliers are a common phenomenon in geographic attributes, typically manifesting as extreme values, bias values, or potential errors (Nirel *et al.* 1998, Simões and Peterson 2018, Tang *et al.* 2024). These outliers disrupt the assumptions of many spatial models, such as regression (both linear and nonlinear), kriging, and geostatistical models, which rely on the regularity of spatial data (Berke 2001, Militino *et al.* 2006, Kim *et al.* 2016). The presence of spatial outliers complicates predictions makes the interpretation of spatial models less reliable. Therefore, outliers are crucial for spatial prediction because they represent significant deviations from expected patterns in spatial data, often indicating critical events or phenomena (Sun *et al.* 2019, Zhang and Yang 2019). Identifying and accounting for spatial outliers can help detect extreme weather events (Kim *et al.* 2024), disease outbreaks (Yang *et al.* 2020), or anomalies in traffic patterns (Yujun *et al.* 2019). Unlike traditional outliers, which deviate globally, spatial outliers are determined through local comparisons with neighboring

data points. Proper handling of these outliers is essential to improve prediction accuracy, as misidentification or exclusion can lead to biased results (Araki *et al.* 2017, Baba *et al.* 2022). Advanced techniques like median kriging with robust estimators allow for the integration of outliers into models while minimizing their negative impact, enhancing the robustness of spatial predictions (Sun *et al.* 2019).

Existing methods for quantifying and detecting outliers that might be helpful for spatial prediction can be classified into the following categories. One common approach is based on spatial cross-outliers, which simultaneously detect outliers by analyzing multiple types of spatial point events, utilizing methods like the cross K-function and Delaunay triangulation to establish relationships between points (Shi *et al.* 2018). For genetic data, the Moran spectral outlier detection (MSOD) technique uses spatial eigenvectors to identify outliers based on unusual patterns in the distribution of genetic variation (Wagner *et al.* 2017). Bayesian neural networks (BNNs) have also been applied to outlier detection by quantifying different types of uncertainty, allowing for identifying outlier observations (Liu *et al.* 2010). Another method, bipartite spatial point estimation, leverages Z-scores to detect outliers by measuring deviations between estimated and true spatial point values (Wei *et al.* 2004). For categorical spatial data, the Pair Correlation Ratio (PCR) metric calculates the co-occurrence frequencies of categories across distances, helping identify outliers based on spatial dependency (Liu *et al.* 2014). These diverse methods address the unique complexities of spatial data, providing robust tools for outlier detection in various applications.

However, there are still limitations in the existing studies for spatial prediction. First, existing studies often ignore the role of outliers in enhancing spatial prediction accuracy, which leads to prediction bias, information loss, and the misrepresentation of spatial variability. Second, there needs to be more research focused on quantifying local outlier strengths and assessing their impact on spatial predictions, leading to a gap in understanding their true significance. Third, various models fail to utilize data from unsampled locations, which could refine predictions by incorporating information beyond sample areas. Predictions without considering unsampled locations information in underestimating extreme values, both high and low, further limit the effectiveness of current spatial prediction models.

This study develops a second-dimension outliers (SDO) model for more accurate spatial prediction through effective outlier incorporation, quantifying local outlier strength, and examining outlier information outside sample locations of the dependent variable (i.e. unsampled locations). An index is then developed to quantify the strength of these outliers, which is used in statistical modeling to assess the relationship between dependent variables and the second-dimension outlier strength. The SDO model is validated by comparing its accuracy with seven existing machine learning models to assess its improvement in spatial prediction. Finally, the developed model is applied to predict wheat production in Australia, enhancing the accuracy of local wheat production predictions.

2. Second-dimension outliers

2.1. Concept of the second-dimension outliers (SDO)

Existing spatial prediction models primarily rely on information derived from sample points, employing machine learning or regression techniques to predict outcomes.

However, these methods often ignore the influence of data at surrounding unsampled locations or data outside sampling locations, which may provide valuable information for predicting the dependent variable. To address this limitation, we developed a novel approach called second-dimension outliers (SDO). The SDO method identifies spatial outliers within a defined local buffer surrounding each sample point, providing additional contextual information. Then, we generate a set of SDO variables, which capture the outlier information from neighboring points. When combined with the original sample point data, these SDO variables serve as new explanatory variables, facilitating the construction of more accurate prediction models.

The SDO method provides several key advantages over traditional aspatial models. Incorporating local outlier information accounts for spatial heterogeneity that could otherwise be missed. Employing outliers can improve prediction accuracy with smoother results, more precise maximum and minimum values, and smaller prediction errors. In addition, the SDO model facilitates downscaling, allowing for high-resolution predictions from coarse observational data. The results in significantly enhanced spatial resolution and more detailed predictive surfaces are critical for environmental monitoring and agricultural forecasting applications.

2.2. SDO model

This study developed an SDO model for examining the relationship between spatial dependent variables and the local outliers based on second-dimension outlier strength. We denote the sample points of the sampled location as u and the unsampled grid points as v. In this study, the unsampled locations are defined as the locations and regions outside the sample locations of the dependent variable. The construction of the SDO model involves four main steps (Figure 1).

First, we determine a set of local ranges (i.e. sizes of buffers) around each sample point, indicating that the outlier information of the explanatory variable will be calculated at different spatial ranges. The selection of buffer sizes was based on the spatial characteristics, scales, and attributes of the dependent variable of the study area. Larger buffer sizes capture more spatial outlier information, and the outliers they introduce can influence the central sampling point, reducing its estimation accuracy and affecting the overall prediction precision. Therefore, buffer sizes were carefully chosen to align with the spatial scale of the study area.

In general, the number of buffers should be selected within the range of 5 to 10 to balance computational efficiency and statistical reliability, avoiding excessively small sample sizes or overly large data volumes. The buffer threshold is typically determined as 10% to 20% of the maximum pairwise distance between sampling points, ensuring an appropriate spatial range for capturing outlier information (Shi *et al.* 2021, Zhang *et al.* 2021, Qi *et al.* 2022). Following these criteria, this study adopted six buffer sizes (ranging from 2 to 7) for the simulation dataset and seven buffer sizes (ranging from 100 km to 700 km) for the case study dataset. The following formula defines the specific buffer thresholds and intervals used in this study:

$$b_\alpha, \alpha = 1, 2, \ldots, m \tag{1}$$

where b_α represents the size of the α-th buffer, m is the buffer threshold.

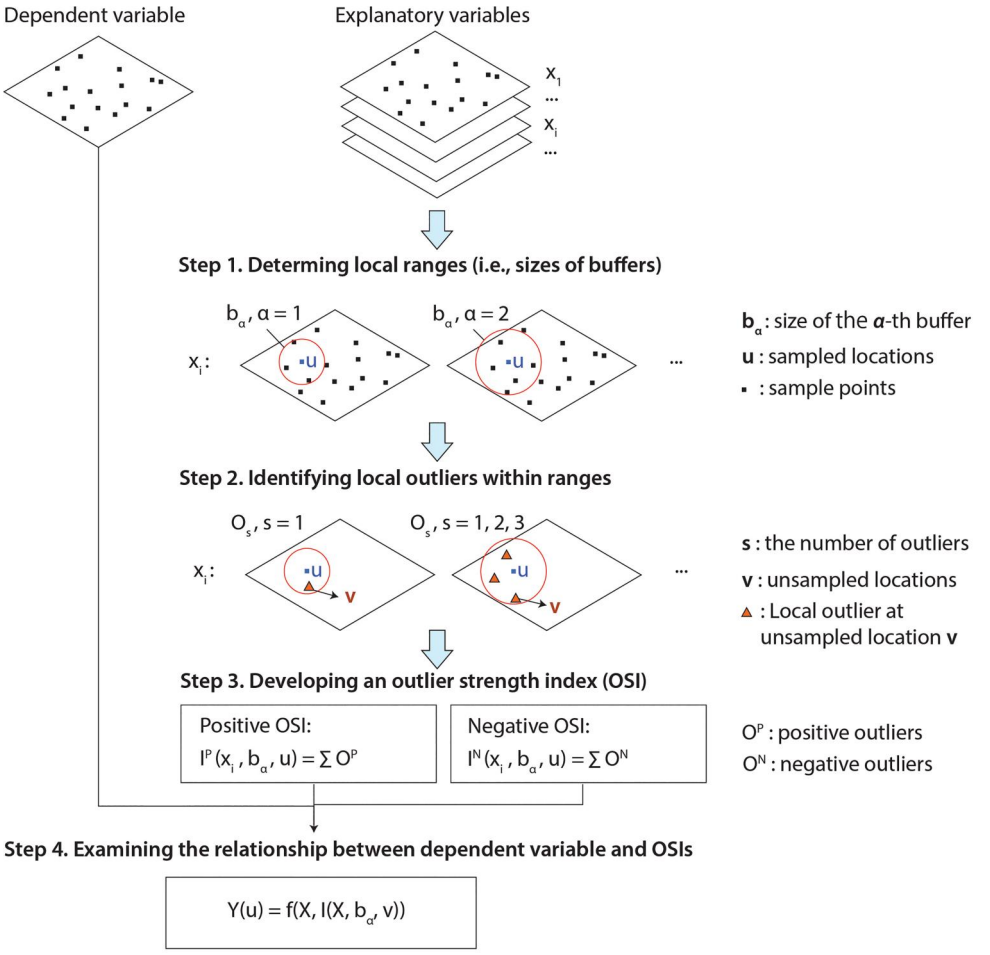

Figure 1. Schematic overview of second-dimension outliers (SDO) model for examining the relationship between dependent variables and spatial local outliers.

The second step is to identify local outliers within ranges. For each buffer range, local outliers are identified by comparing the values of surrounding points (within the buffer) to the sample point. The outliers in this study are defined based on their deviation from the expected value, with positive outliers (O_s^P) exceeding the mean plus two standard deviations ($\overline{x} + 2\sigma$) and negative outliers (O_s^N) falling below the mean minus two standard deviations ($\overline{x} - 2\sigma$). For an explanatory variable, the local outliers for a location u are presented as:

$$O_s(u), s = 1, 2, \ldots, n \tag{2}$$

where $O_s(u)$ is a vector representing the values of outliers at unsampled locations v within the buffer area surrounding location u, where s indexes the outliers from 1 to n, and n represents the total number of outliers within the given buffer.

The third step is to develop the outlier strength index (OSI), which quantifies the cumulative magnitude of identified outliers (positive or negative) for each explanatory variable. The OSI is calculated as the sum of outlier values extracted from multiple buffer sizes, thereby incorporating spatial outliers across different spatial scales. The following equation illustrates how OSIs are computed to generate (SDO) variables, which capture multiscale spatial outlier information.

$$\begin{aligned} I^P(X, b_\alpha, v) &= \sum_{s=1}^{m} O_s^P(b_\alpha, v) \\ I^N(X, b_\alpha, v) &= \sum_{t=1}^{n} O_t^N(b_\alpha, v) \\ I(X, b_\alpha, v) &= \left[I^P(X, b_\alpha, v),\ I^N(X, b_\alpha, v)\right] \end{aligned} \tag{3}$$

where X represents the spatial explanatory variable, b_α is the α-th buffer distance, and v denotes an unsampled location within buffer b_α surrounding the target sample location u. $O_s^P(b_\alpha, v)$ and $O_t^N(b_\alpha, v)$ are the identified positive and negative outlier values (based on a z-score threshold) within the buffer area. I^P and I^N represent the cumulative positive and negative outlier intensities at a specific scale. Each buffer yields a pair of values, I^P and I^N, for each explanatory variable, forming a set of SDO variables. These variables are incorporated into the prediction model as multi-scale spatial features, enabling the model to account for spatial outliers at different distances and improve prediction accuracy.

Finally, we examine the relationship between the dependent variable and outlier strength indexes (OSIs). The original explanatory variables from the sample points are combined with the SDO variables (i.e. I^P and I^N) to train the prediction model. Machine learning algorithms such as random forest (RF) and support vector machines (SVM) are used to build the models, then evaluate and compare the accuracy of the aspatial and SDO models.

$$Y(u) = f(X, I(X, b_\alpha, v)) \tag{4}$$

where $Y(u)$ is the spatial dependent variable at the sampled location u. X represents the original explanatory variables, which are observed at sampled locations. $I(X, b_\alpha, v)$ denotes the second-dimension outlier variables derived from the outlier strength of X at unsampled locations v within different buffer sizes b_α. The function $f(\cdot)$ represents the prediction model (e.g. SVM, RF) trained using both X and the derived outlier variables. It is important to note that no outliers are removed from X; rather, outlier information is extracted and quantified to enhance spatial prediction.

2.3. Simulation data analysis

A set of simulated data was first employed to evaluate the performance of the proposed SDO model. The observation data (sample points) were randomly generated within a 20 $\times$ 20 region, with random longitude and latitude values. The dependent variable y, and four explanatory variables, x_1, x_2, x_3, and x_4, were also generated randomly. Figure 2 presents the spatial distribution of the dependent variable (y) and the

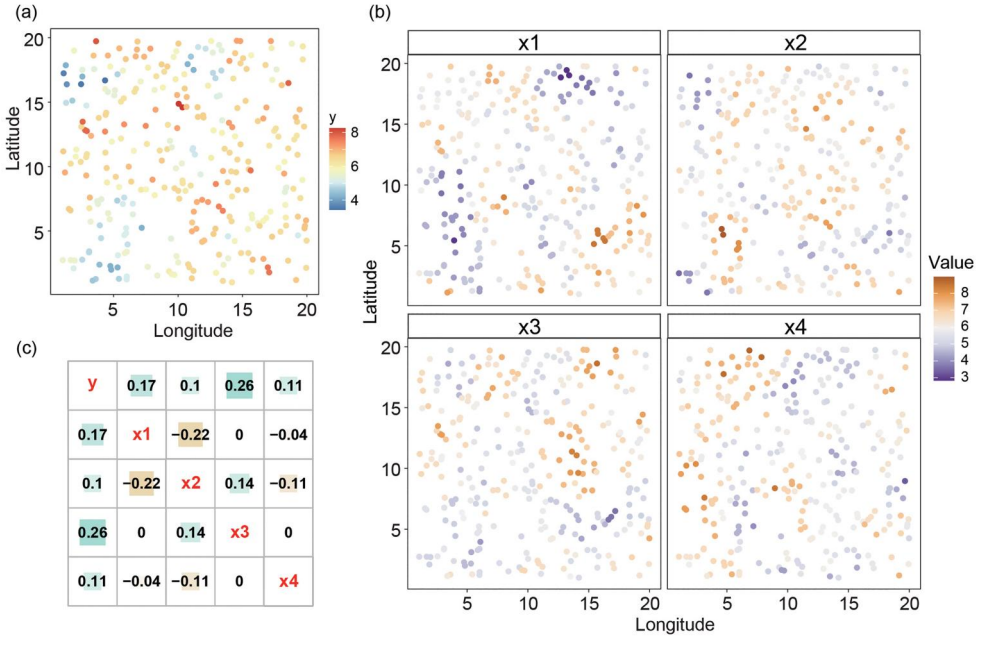

Figure 2. Summary of simulation data including dependent variable (a), explanatory variables (b), and the correlation analysis with them.

four explanatory variables, along with the results of a correlation analysis among them.

The sample point dataset consists of longitude (*lon*), latitude (*lat*), *y*, and the four explanatory variables (x_1, x_2, x_3, and x_4). The aspatial prediction model is designed to estimate *y* based on these explanatory variables. The dataset includes a total of 300 sample points, with *y* values ranging from 3.42 to 8.26 and an average value of 6.09. The target prediction grid follows a standard 20 × 20 structure, where longitude and latitude values range from 1 to 20. Each grid point contains the same four explanatory variables (x_1, x_2, x_3, and x_4), providing a consistent spatial framework for prediction. This setup ensures that the model can be systematically evaluated while balancing training data availability and predictive robustness. We employed five-fold cross-validation to ensure a rigorous evaluation, using 80% of the dataset (240 sample points) for training and 20% (60 sample points) for validation in each fold.

For the SDO model, we used the locations of sample points and grid points to identify positive and negative outliers of explanatory variables within different buffer sizes. These outlier data were then used to create the outlier strength index (OSI) and generate SDO variables. Figure 3 illustrates the positive and negative SDO variables generated for the explanatory variables x_1, x_2, x_3, and x_4 at buffer sizes 2 through 7, based on the spatial locations of the sample points (resulting in a total of 24 SDO variables). These SDO variables, combined with the original explanatory variables, were used to build and train the predictive models. Machine learning algorithms, particularly random forest (RF) and support vector machines (SVM) were applied to compare the results of aspatial machine learning models (aspatial RF and SVM) with SDO-integrated models (SDO-RF and SDO-SVM).

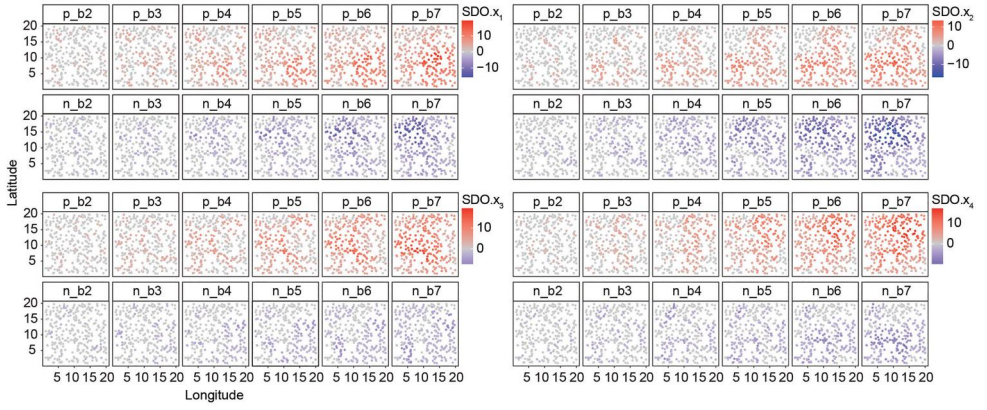

Figure 3. The second-dimension variables of the explanatory variables of the simulated data are generated by the SDO model. The letters p and n mean positive and negative outliers, respectively. The letters b2, b3, …, b7 mean the buffers with sizes varying from 2 to 7.

Figure 4(a,b) show the spatial prediction results of SDO-RF, aspatial RF, SDO-SVM, and aspatial SVM, along with the Pearson correlation coefficient (R) between the predicted and actual values. The SDO-RF model achieved better predictive performance with a higher R value. The SDO models produced smoother predictions than their aspatial counterparts and were more accurate in capturing spatial outliers (high and low values). Notably, the R value of SDO-SVM increased from 0.48 in the aspatial SVM model to 0.7, representing a 45.8% improvement. In addition, the importance of the variables used in the aspatial RF and SDO-RF models was compared to assess the contribution of each variable to the model's prediction (Figure 4(c)). It was observed that x_3 had the highest contribution in the For the SDO model, we used the locations of sample points and grid points to identify positive and negative outliers of explanatory variables within different buffer sizes. These outlier data were then used to create the outlier strength index (OSI) and generate SDO variables. Figure 3 illustrates the positive and negative SDO variables generated for the explanatory variables x_1, x_2, x_3, and x_4 at buffer sizes 2 through 7, based on the spatial locations of the sample points (resulting in a total of 24 SDO variables). These SDO variables, combined with the original explanatory variables, were used to build and train the predictive models. Random forest (RF) and support vector machines (SVM) machine learning algorithms were applied to compare the results of aspatial machine learning models (aspatial RF and SVM) with SDO-integrated models (SDO-RF and SDO-SVM). In both the aspatial RF and SDO-RF models, x_3 was the most influential variable. In the SDO-RF model, this was further reinforced by four of the top ten contributing variables being either x_3 itself or its SDO-derived variables, highlighting the role of spatial outlier information in improving predictive accuracy.

The results of the above simulation case indicate that the SDO model provides accurate predictions in regression based on local spatial outliers. SDO model outperforms aspatial models in prediction accuracy, smoothness of results, outlier detection, and overall model fit.

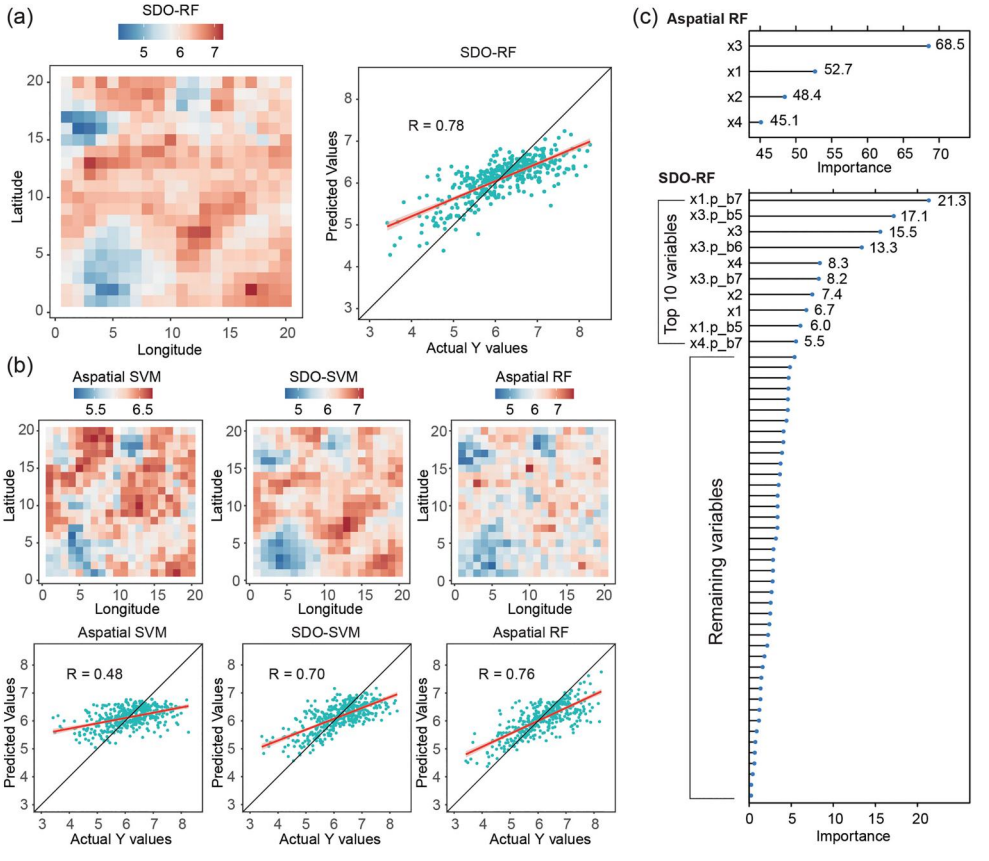

Figure 4. Results of SDO modeling on simulation data using random forest (RF) and support vector machine model (SVM). (a) SDO prediction result using RF, (b) aspatial prediction results based RF and SVM, and SDO-SVM prediction, (c) Importance of the individual variables for the aspatial (upper) and SDO (lower) RF models.

3. Case study: predicting wheat production in Australia using SDO model

3.1. Study area and data

In this study, we applied the SDO model to the downscaling prediction of wheat production in Australia, focusing on 179 local government area (LGA) regions within the Australian wheat belt (Feng *et al.* 2022) for 2021. Wheat production data for these LGAs were obtained from the Australian Bureau of Statistics (ABS) (Australian Bureau of Statistics 2021). Specifically, our predictive models used the average wheat production for each LGA as the dependent variable. Figure 5 illustrates the distribution of average wheat production across these 179 LGAs in the Australian wheat belt for 2021, highlighting the major wheat-producing areas in central New South Wales, southern Victoria, and western Australia.

For the explanatory variables influencing wheat production, we selected three main categories of data that impact wheat production significantly: climate variables, environmental data, and soil properties, as detailed in Table 1. The climate variables include air temperature and total precipitation, sourced from the ERA5-Land dataset,

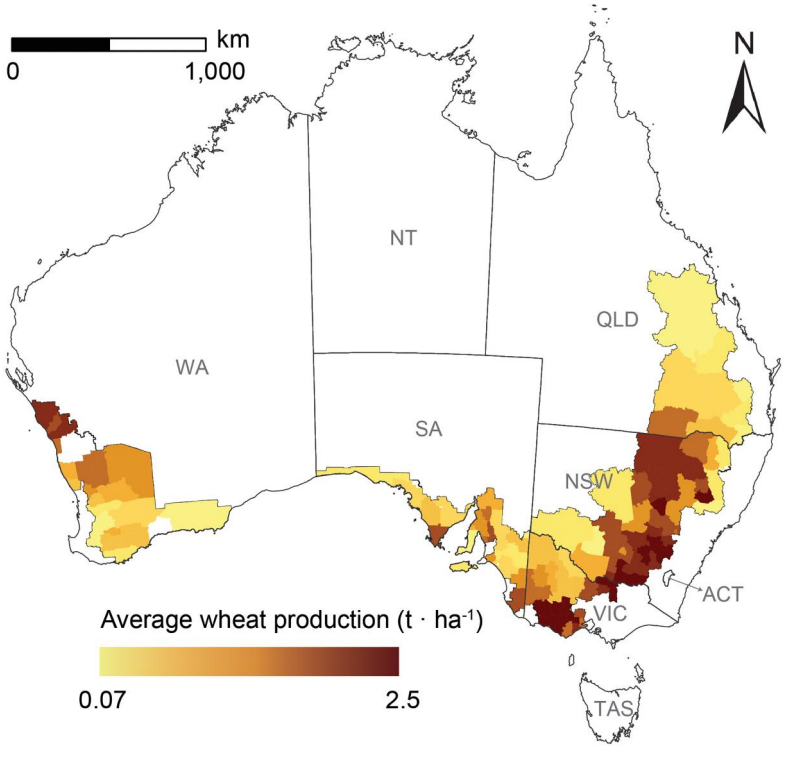

Figure 5. Annal wheat production in Australia's LGAs within the wheat belt in 2021.

Table 1. A summary of explanatory variables that potentially affect spatial disparities of wheat production.

Category	Variable	Code	Product	Resolution
Climate	Air temperature	AT	ERA5_land	0.25°
	Total precipitation	TP	ERA5_land	0.25°
Environment	Evapotranspiration	ETa	CMRSET Landsat V2.2	30 m
	Normalized difference vegetation index	NDVI	MOD13A2 V6.1	1000 m
	Enhanced vegetation index	EVI	MOD13A2 V6.1	1000 m
Soil data	Total Nitrogen	NTO	CSIRO/SLGA	92.77 m
	Total Phosphorus	PTO	CSIRO/SLGA	92.77 m
	Sand	SND	CSIRO/SLGA	92.77 m
	Silt	SLT	CSIRO/SLGA	92.77 m

which provides mean values for 2021 across the wheat belt at a spatial resolution of 0.25° (Hersbach *et al.* 2023). Environmental variables include actual evapotranspiration (ETa), normalized difference vegetation index (NDVI), and enhanced vegetation index (EVI). The ETa data were derived from the CMRSET Landsat V2.2 dataset available on the Google Earth Engine (GEE) platform, with a spatial resolution of 30 meters (Guerschman *et al.* 2022). NDVI and EVI data were obtained from the MOD13A2 V6.1 dataset, also available on GEE, with a spatial resolution of 1000 meters (Didan 2021). Soil data were sourced from the Soil and Landscape Grid of Australia (SLGA) dataset produced by CSIRO, with key attributes such as total nitrogen (NTO), total phosphorus (PTO), sand proportion (SND), and slit proportion (SLT) selected for their relevance to

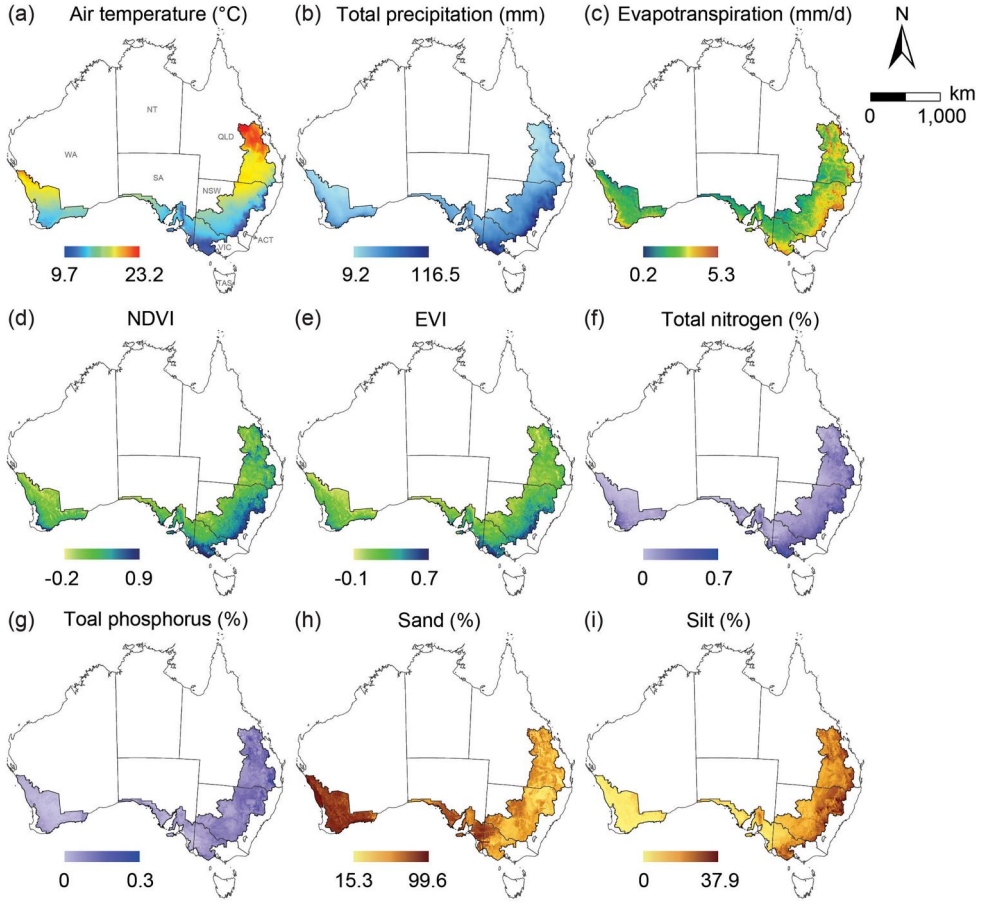

Figure 6. Spatial distributions of explanatory variables in 2021. (a) and (b) Climate variables, (c)–(e) Environmental variables, (f)–(i) Soil attributes.

wheat production in Australia (Rossel *et al.* 2015). Figure 6 shows the spatial distribution of explanatory variables within the wheat belt in Australia, including climate variables, environmental variables, and soil attributes.

Because this study considers only LGAs located within the Australian wheat belt, we calculated the mean wheat production for each of those LGAs and masked every environmental layer with the national wheat-crop map before averaging pixel values inside the same boundaries. All spatial datasets were projected using the `GDA_1994_Australia_Albers` (EPSG:3577) coordinate system to ensure consistency in spatial alignment and area-based calculations. For the prediction grid, each explanatory layer was resampled on the GEE platform so that the mean value within every grid cell represents the corresponding variable at that finer scale.

3.2. Experiment design

The experimental design for applying the SDO model to predict wheat production in Australia is as Figure 7 shows. First, wheat production data and explanatory variables

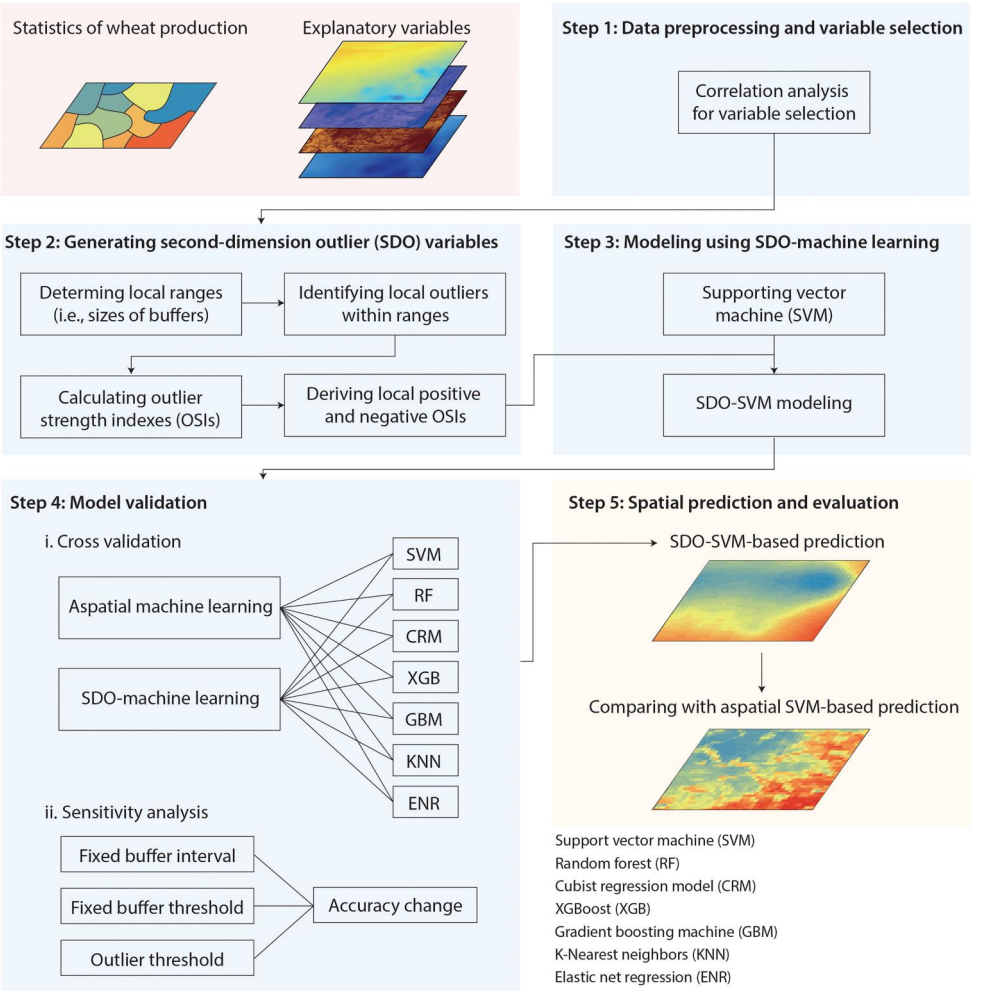

Figure 7. Main steps of the second-dimension outliers (SDO) model for predicting wheat production in Australia.

for the study areas were collected from shapefiles, remote sensing data, and statistical sources, followed by data preprocessing, correlation analysis, and variable selection. Second, spatial outliers were identified at different buffer sizes to generate second-dimension variables for the explanatory variables. Third, the SDO model was integrated with the SVM for modeling. While RF performed better in the simulation, SVM showed higher accuracy and smoother predictions in the case study, and was thus selected for detailed evaluation. Fourth, the SDO was also combined with six other machine learning models, and its statistical performance was compared to non-spatial machine learning models to assess the effectiveness of the SDO model. Finally, the SVM model, which performed best among the seven models, was selected for wheat production predictions. A comparison was made between the results of the SDO-SVM and aspatial SVM models, focusing on variable importance, cross-sectional predictions

for the New South Wales (NSW) region, and the density distribution of predictions across various states.

The first step includes data preprocessing and variable selection through correlation analysis to identify explanatory variables with a significant impact on wheat production. Through correlation analysis, nine explanatory variables with a higher correlation to wheat production were selected. The regional average represented observations for each LGA region, while the grid cells for predictions had a resolution of 10,000 meters.

In the second step, we identify local outliers under different buffer sizes by determining local ranges and calculating the outlier strength indexes (OSIs) to generate the second-dimension outlier variables. Based on the SDO concept and the spatial locations of sample points and grid points, seven buffer sizes were determined, ranging from 100 km to 700 km with intervals of 100 km. Then, local spatial outliers were identified within each buffer range, and outlier strength indexes (OSIs) were calculated to generate corresponding local SDO explanatory variables. The SDO variables were classified into positive SDO and negative SDO variables based on whether they indicated positive or negative outliers. For each variable in this case, 14 positive and negative SDO variables were generated across the seven buffer distances.

Third, we employed SDO-based machine learning for spatial prediction of wheat production. We integrated SDO with support vector machine (SVM) due to SVM having strong generalization ability, robustness to outliers, and high interpretability (Cherkassky and Ma 2004, Yu and Kim 2012, Zhu *et al.* 2024). SVM fits the data by minimizing prediction error while maintaining a margin that ensures most data points fall within it (Yu and Kim 2012). SVM uses a kernel function to map the input data into a higher-dimensional space, allowing the model to find an appropriate hyperplane for regression (Abakar and Yu 2014). In this study, the SVM model employed the Gaussian kernel function (also called radial basis function kernel, RBF) (Wang *et al.* 2004, Ding *et al.* 2021), which is expressed as follows:

$$K(\overrightarrow{x_i}, \overrightarrow{y_i}) = e^{-\gamma \| \overrightarrow{x_i} - \overrightarrow{y_i} \|^2} \tag{5}$$

where $\overrightarrow{x_i}$ and $\overrightarrow{y_i}$ are sample points or vectors, γ defines the influence of individual samples on the overall hyperplane. When γ is small, individual samples have a broader influence on the classification hyperplane, increasing the likelihood of being support vectors. In contrast, a larger γ sharpens the Gaussian function, leading to a more complex model and a higher risk of overfitting. The integration of SDO and SVM enhances robustness to local spatial outliers, improving regression accuracy by optimizing hyperplane selection.

In the fourth step, the prediction results of the SDO model were compared with aspatial models and analyzed for sensitivity to verify the accuracy and robustness of the constructed model. In addition to SVM used in step 3, six other machine learning models were selected for comparative analysis: random forest (RF), cubist regression model (CRM), XGBoost (XGB), gradient boosting machine (GBM), K-nearest neighbors (KNN), and elastic net regression (ENR). Combining each machine learning method with aspatial data formed the aspatial models, referred to as aspatial SVM, RF, CRM, XGB, GBM, KNN, and ENR. When combined with SDO, the models were named SDO-SVM, SDO-RF, SDO-CRM, SDO-XGB, SDO-GBM, SDO-KNN, and SDO-ENR. To ensure a

rigorous evaluation, we employed five-fold cross-validation, where each model was trained on 80% of the data and validated on the remaining 20% in each fold. In addition, a sensitivity analysis was conducted to assess the robustness of the SDO model with respect to the buffer interval, buffer threshold, and standard deviation threshold for outlier identification, ensuring stable performance across these parameter configurations. The modeling accuracy and errors of the aspatial and SDO models were compared using R^2, root mean squared error (*RMSE*), and mean absolute error (*MAE*) as evaluation metrics. The formulas for these metrics are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{8}$$

where y_i denotes observed values, $\hat{y}_i$ indicates the predicted values, $\overline{y}$ is the mean of the observed values, and n represents the number of observations.

Finally, we predicted the spatial distribution of wheat production across Australia's wheat belt using the SDO-SVM model and evaluated the prediction results. We selected the SVM model, which performed best in the previous step and integrated it with the SDO model to compare its predictions with those of the aspatial SVM model. We compared the R^2, *RMSE*, and *MAE* metrics for the SDO-SVM and aspatial SVM models. Additionally, we analyzed cross-sectional predictions within the NSW wheat belt and compared the density distribution characteristics of the prediction results for each state, further demonstrating the accuracy and effectiveness of the SDO model developed in this study.

4. Results

4.1. Data preprocessing and variable selection

The SDO model developed in this study supports modeling and training of data at different scales. Data preprocessing includes handling missing data, generating a 10,000 m resolution grid, and resampling the explanatory variable data on each grid cell.

Correlation analysis was conducted for all variables, with the results presented in Figure 8. Significant correlations were observed between explanatory variables and the dependent variable (wheat production), particularly among temperature, evapotranspiration, EVI, NDVI, total precipitation, silt, and sand. Initially, wind speed was considered a potential explanatory variable; however, the correlation analysis revealed a very low correlation between wind speed, wheat production, and other variables, leading to its exclusion from the final model.

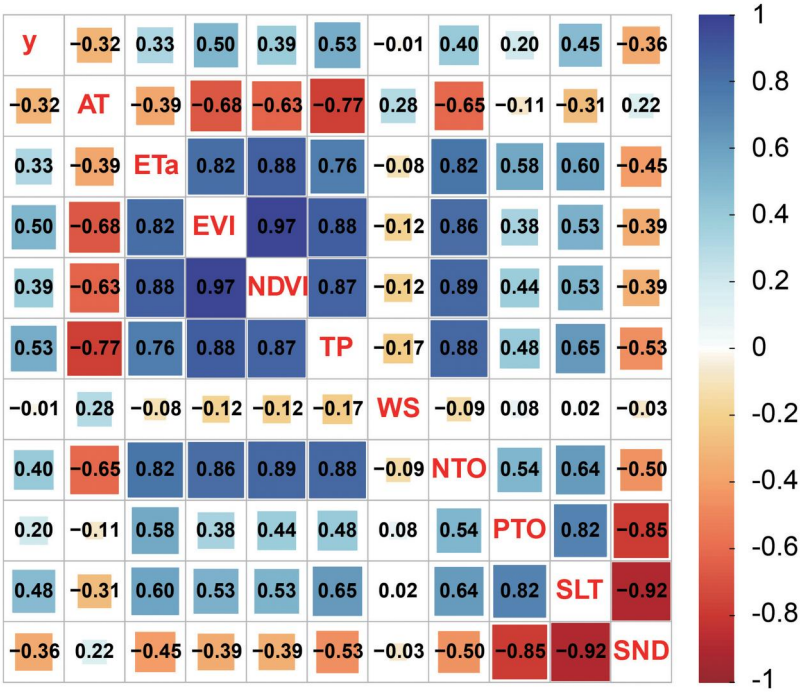

Figure 8. Correlation analysis for wheat production and explanatory variables, including air temperature, evapotranspiration, EVI, NDVI, total precipitation, total nitrogen, total phosphorus, silt, sand, and wind speed.

4.2. Generate SDO variables and modeling using SDO-machine learning

Based on the SDO model and the spatial locations of sample points and grid cells, spatial outliers of explanatory variables were identified across different buffer sizes in the study area, and the corresponding outlier strength index (OSI) was calculated to generate SDO variables. These SDO variables were categorized into positive and negative values: positive variables indicated positive outliers identified by the model, while negative variables represented negative outliers. Figure 9 illustrates the second-dimension outlier variables generated by the SDO model for air temperature across different buffer sizes. As shown in Figure 9, the outlier data for air temperature is primarily concentrated in the eastern regions of Australia, covering parts of New South Wales (NSW), Victoria (VIC), and Queensland (QLD). This phenomenon can be attributed to the region's complex mountainous terrain and its exposure to southeastern trade winds and ocean currents.

Figure 10 presents the spatial distributions of second-dimension outlier variables for total precipitation (TP), normalized difference vegetation index (NDVI), and sand (SND) at buffer distances 100 km, 400 km, and 700 km, selected to illustrate representative multi-scale spatial patterns. As the buffer size increases, the intensity of both positive and negative outlier values become more pronounced, reflecting the model's ability to capture broader-scale outliers beyond local variation. Similar to previous results as Figure 9, this figure also captures both positive and negative outliers generated for different buffer sizes. The intensity of these outliers increases with larger

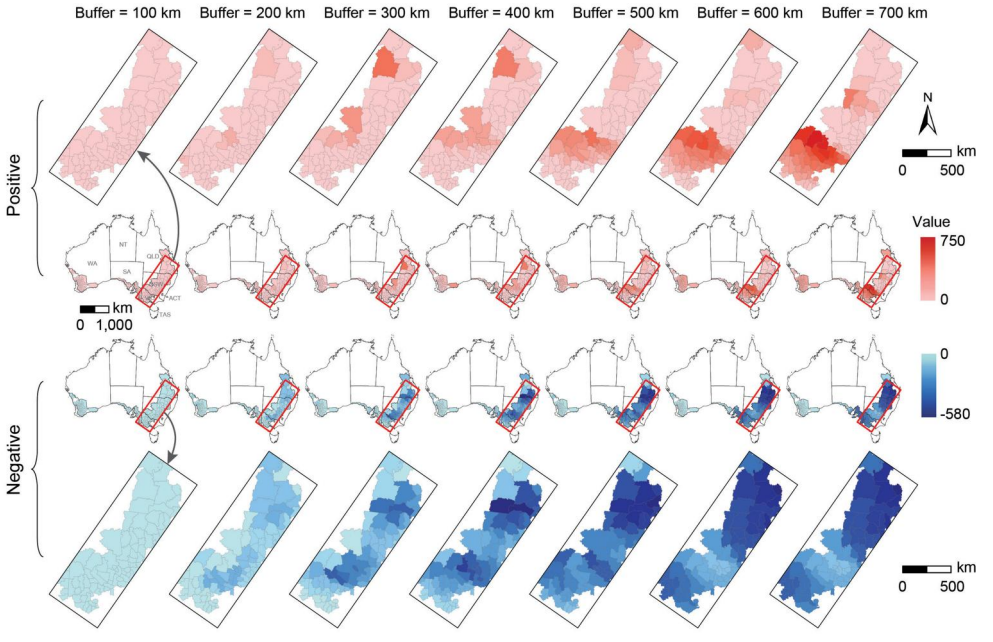

Figure 9. The distribution of the second-dimension temperature variable generated by the SDO model under different buffers, which includes positive outliers (red) and negative outliers (blue).

buffer sizes, indicating a stronger deviation from local spatial patterns as the buffer expands. Then, we combined the SDO variables with the original explanatory variables and employed a support vector machine (SVM) for modeling and training. The preliminary results confirm the effectiveness of incorporating the SDO model, as it improves the model's ability to account for spatial outliers and enhances overall prediction accuracy.

4.3. SDO model validation and sensitivity analysis

The SDO model was integrated with seven machine learning algorithms to construct a series of SDO-enhanced models. Table 2 compares the predictive performance of these SDO models with their aspatial counterparts in terms of R^2, *RMSE*, and *MAE*. The results indicate that most SDO-based models achieved higher R^2 values and lower prediction errors, confirming the benefits of incorporating spatial dependence optimization. In particular, the SDO-SVM model achieved the best overall performance, with the highest R^2 (0.671) and the lowest RMSE (0.311) and MAE (0.231), followed by the SDO-KNN and SDO-XGB models. Although a few models (e.g. CRM and ENR) exhibited slight declines in performance, these variations likely stem from their relatively rigid functional forms or weaker adaptability to spatially dependent features. Overall, the integration of the SDO framework effectively enhances model accuracy by capturing spatially heterogeneous patterns that are often not accounted for by aspatial machine learning approaches.

To assess the robustness of the SDO model, a sensitivity analysis was performed by varying the buffer threshold, buffer interval, and outlier threshold (standard deviation,

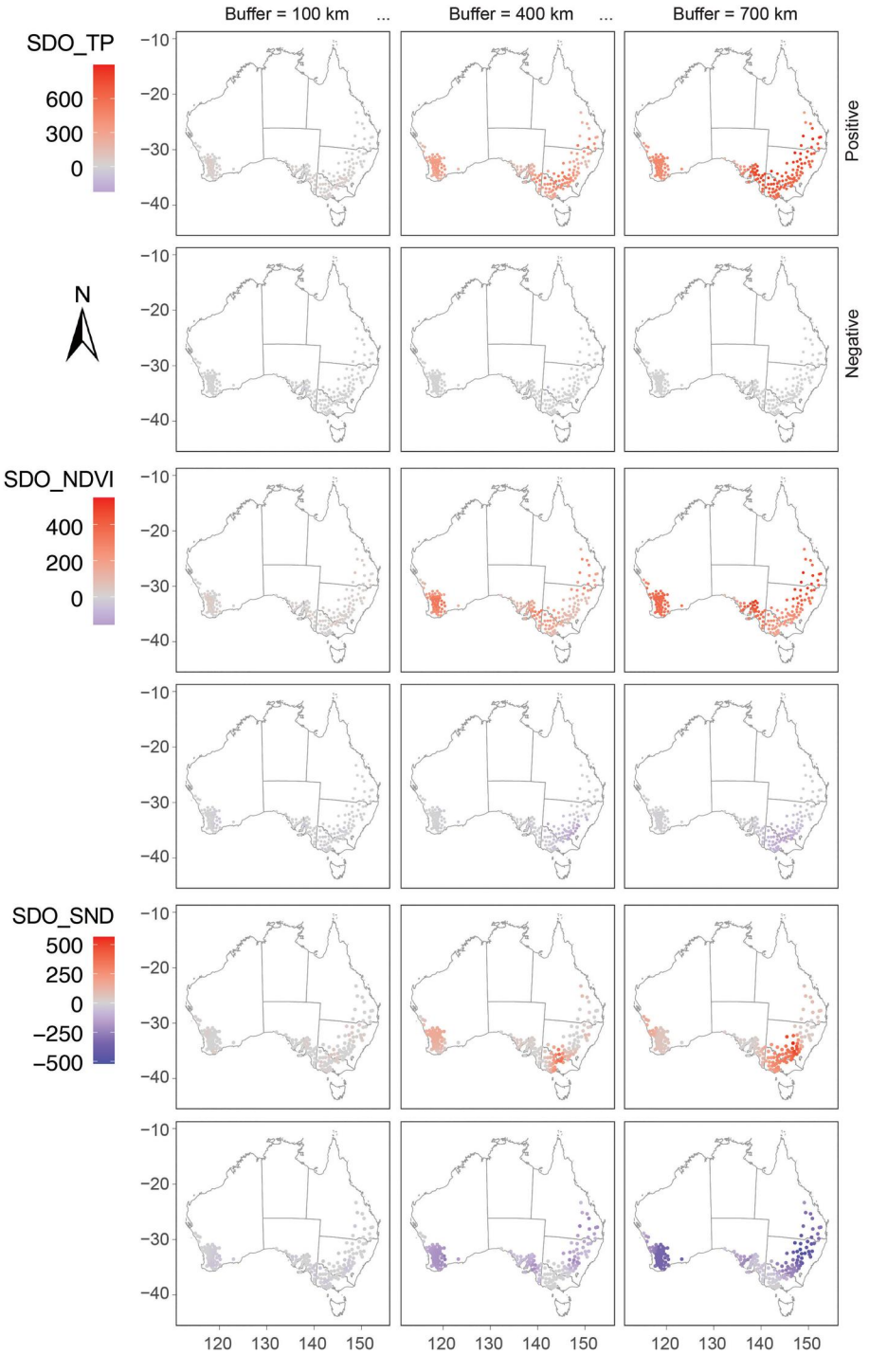

Figure 10. Spatial distributions of second-dimension outlier variables for total precipitation (TP), normalized difference vegetation index (NDVI), and sand (SND) at buffer distances 100, 400, and 700 km, generated using the SDO model.

Table 2. Improvements of model accuracy in machine learning by SDO models compared with aspatial models (bold values indicating the best-performing SDO results).

Model		RF	CRM	XGB	SVM	GBM	KNN	ENR
R^2	Aspatial	0.527	0.634	0.548	0.555	0.496	0.455	0.580
	SDO	0.615	0.590	0.628	**0.671**	0.596	0.637	0.585
	Improvement	16.7%	−6.9%	14.6%	20.9%	20.2%	**40.0**%	0.9%
RMSE	Aspatial	0.368	0.321	0.370	0.355	0.377	0.400	0.343
	SDO	0.338	0.350	0.330	**0.311**	0.337	0.321	0.360
	Reduction	8.2%	−9.0%	10.8%	12.4%	10.6%	**19.8**%	−5.0%
MAE	Aspatial	0.279	0.241	0.286	0.267	0.283	0.310	0.272
	SDO	0.256	0.254	0.243	**0.231**	0.248	0.241	0.265
	Reduction	8.2%	−5.4%	15.0%	13.5%	12.4%	**22.3**%	2.6%

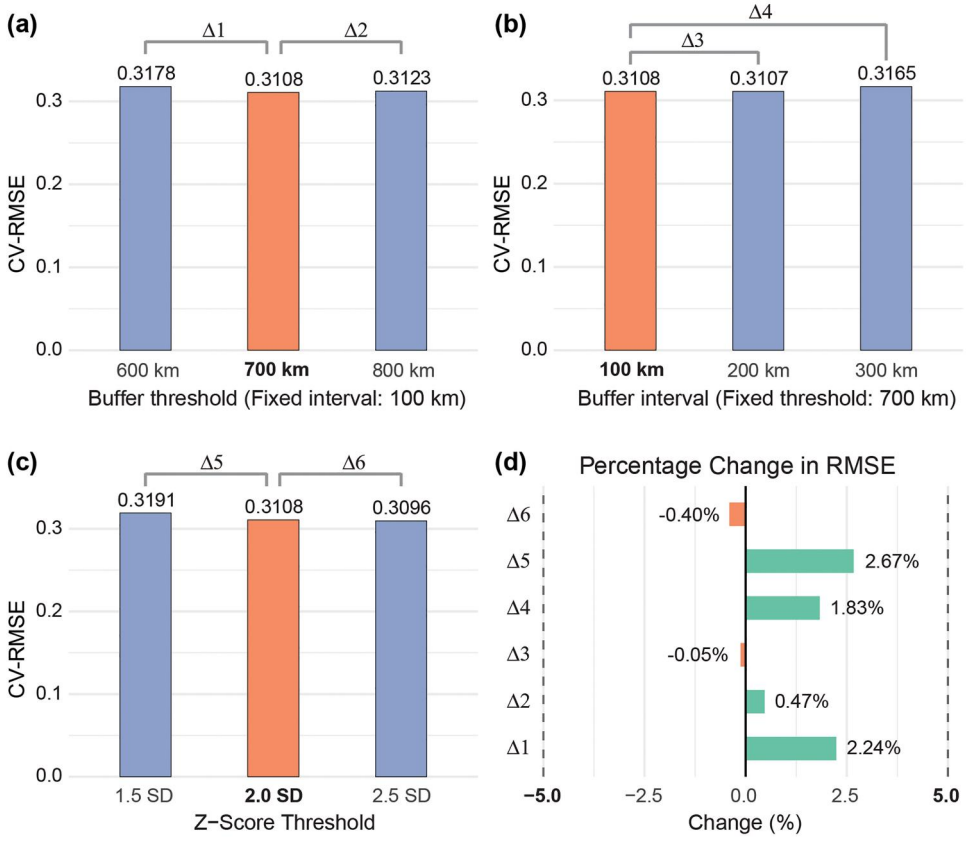

Figure 11. Sensitivity analysis of the SDO model: impact of buffer threshold, buffer interval, and outlier (standard deviation) threshold settings on model accuracy (cross-validation root mean square error, CV-RMSE).

SD) parameters. Figure 11 presents the cross-validated root mean square error (CV-RMSE) under different parameter settings. In Figure 11(a), the buffer interval was fixed at 100 km, and the buffer threshold was varied among 600 km, 700 km, and 800 km. The CV-RMSE remained generally stable, with the lowest value (0.3108) observed at 700 km, relative to 700 km, CV-RMSE is higher by 2.24% at 600 km and by 0.47% at 800 km. In Figure 11(b), when the buffer threshold was fixed at 700 km, the CV-RMSE

slightly decreased (−0.05%) when the buffer interval increased from 100 km to 200 km, but rose by 1.83% at 300 km. In Figure 11(c), the sensitivity of the model to the selection of the outlier identification threshold was examined by testing 1.5, 2.0, and 2.5 standard deviations (SD). The CV-RMSE ranged narrowly from 0.3096 to 0.3191, with the minimum error again occurring at 2.0 SD, indicating that moderate thresholds for identifying outliers yield more stable model performance. Figure 11(d) further illustrates the percentage changes in CV-RMSE under different parameter settings, with all variations remaining within a ± 5% range. These findings confirm the robustness of the SDO model, indicating that its predictive performance remains stable across different spatial parameter configurations. The results demonstrate integrating second-dimension outlier variables enhances model generalizability without being overly sensitive to specific buffer parameter selections.

4.4. Spatial prediction and evaluation

Figure 12 compares the aspatial SVM (a) and SDO-SVM (c) prediction maps. The SDO-SVM results exhibit smoother and more spatially continuous patterns while preserving local variability, whereas the aspatial model shows more fragmented and uneven distributions. The variable importance analysis (b, d) indicates that precipitation (TP), enhanced vegetation index (EVI), and air temperature (AT) remain the most influential predictors in both models. In the SDO-SVM model, several second-dimension outlier variables, particularly those related to NDVI and precipitation within 300–700 km buffer ranges, as well as sand content (SND), rank among the top contributing factors. This finding suggests that incorporating contextual outlier information enhances the model's ability to capture fine-scale spatial heterogeneity in wheat production.

To further evaluate the spatial performance of the SDO model, two cross-sections were analyzed within the wheat belt region of New South Wales (Figure 13). Section A (southwest–northeast) and Section B (northwest–southeast) were selected to represent typical spatial gradients of wheat production. Figure 13(b,c) compare the predicted wheat production profiles from the SDO-SVM and aspatial SVM models along these sections. While both models reproduce the general spatial trends, the SDO-SVM predictions exhibit smoother transitions and better alignment with observed patterns, particularly in regions of rapid yield variation. Compared with the aspatial SVM, which shows sharp local oscillations, the SDO-SVM more effectively represents gradual changes and preserves regional continuity. The spatial maps in Figure 13(d–f) further confirm that the SDO-SVM yields a more coherent distribution of high and low yield zones, demonstrating improved spatial consistency while maintaining predictive accuracy.

Figure 14 compares the predicted average wheat production density distributions from the SDO-SVM (red) and aspatial SVM (blue) across Australia and its individual state. Nationally, the two models show similar central tendencies (dashed lines) and overall shapes. Differences concentrate in the extremes, the green shading highlights the top 5% and bottom 5% ranges captured by SDO, where the red and blue curves diverge most. In several states (e.g. NSW, VIC, WA) the SDO curve is elevated over parts of one or both tails, whereas in others (e.g. QLD, SA) the advantage is mixed

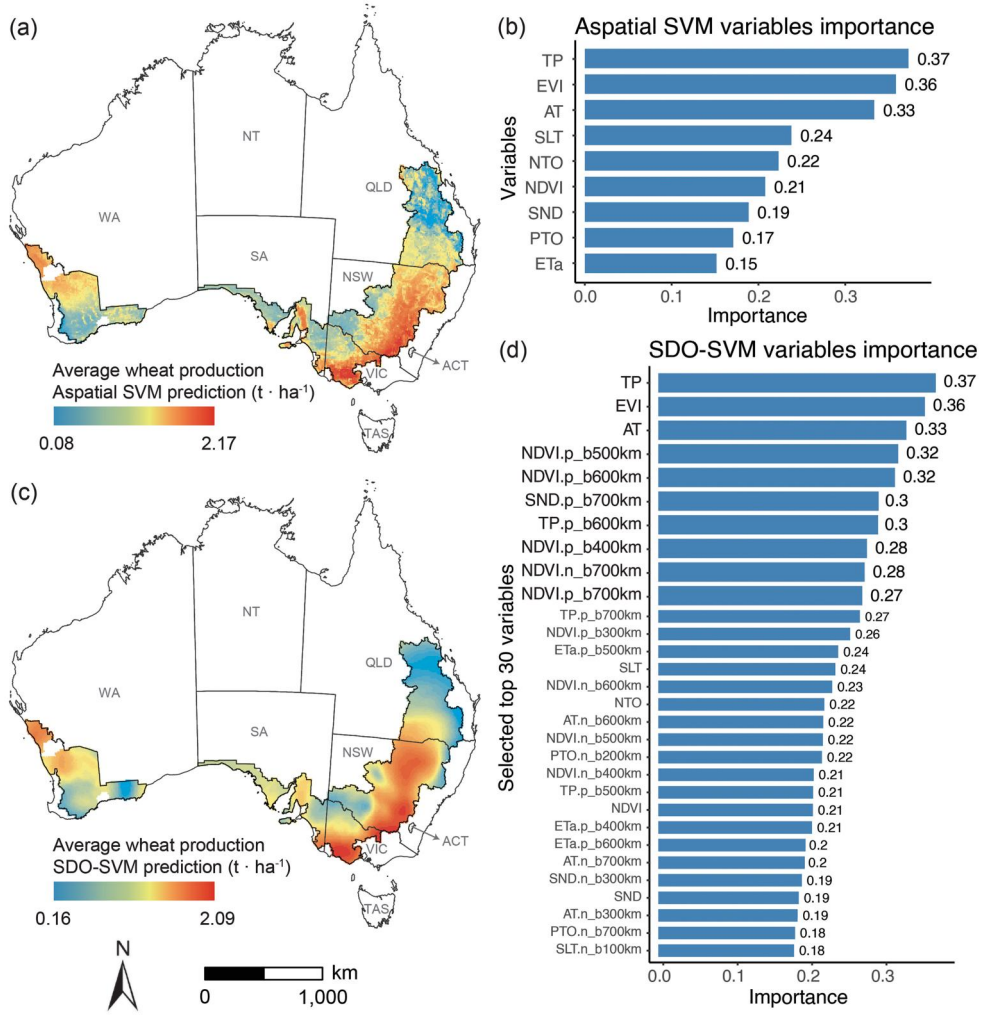

Figure 12. Prediction of average wheat production in Australia based aspatial model (a) and SDO machine learning model (c). The importance of top 30 SDO variables (d) and aspatial variables (b) in the machine learning model.

and the blue curve exceeds red over some tail segments. Since maximum and minimum ranges are where outlier information is most likely to concentrate, the shaded regions indicate that SDO tends to capture a broader share of extreme values while keeping the mean largely unchanged.

5. Discussion

5.1. Predictive benefits of second-dimension outliers

This study introduces the concept of second-dimension outliers (SDO) and develops a series of SDO-based machine learning models for spatial prediction. These models enhance prediction accuracy and reduce model error by incorporating outlier information beyond sample locations. The SDO model establishes relationships between the

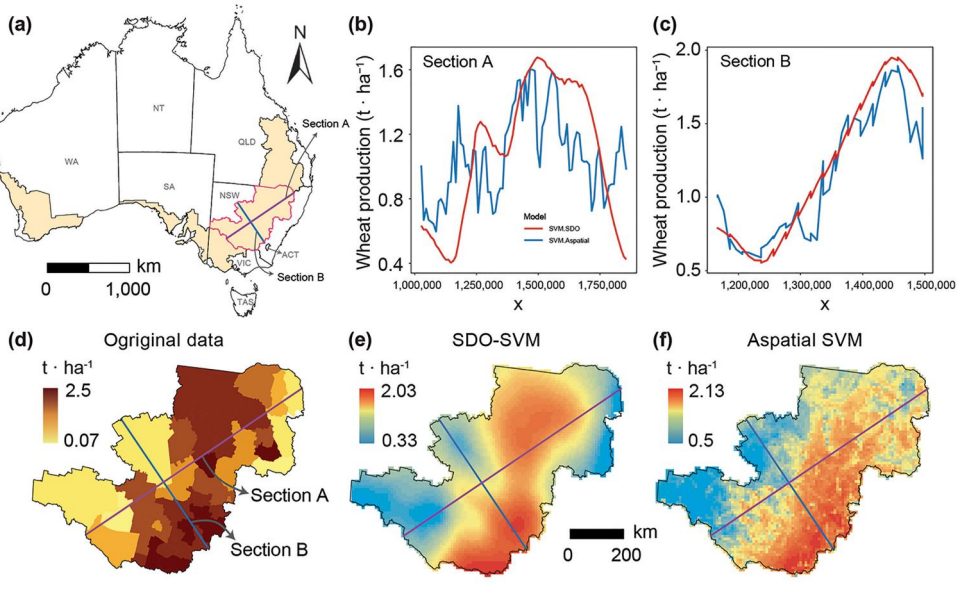

Figure 13. Comparison of SDO and aspatial model predictions along cross-sections in the wheat belt region of New South Wales. (a) Locations of section A (southwest to northeast) and section B (northwest to southeast) in the wheat belt region of New South Wales. (b, c) Predicted wheat production along sections A and B, comparing SDO-SVM and aspatial SVM models. (d) Original wheat production data with cross-sections overlaid. (e, f) Predicted wheat production from the SDO-SVM model (e) and the aspatial SVM model (f), with overlaid cross-sections.

dependent variable and the outlier characteristics of predictor variables at locations outside the sample points, allowing for a more comprehensive representation of spatial outliers. By generating localized SDO variables across multiple buffer distances, the model captures complex geographical and environmental patterns that conventional aspatial models often overlook.

Spatial outliers are crucial in improving prediction accuracy by capturing outliers spatial patterns, explaining local extremes, mitigating errors from spatial heterogeneity, and enhancing model robustness (Lu *et al.* 2009, Baba *et al.* 2022, Hu *et al.* 2025). Spatial autocorrelation suggests that observations within neighboring regions tend to be similar (Overmars *et al.* 2003, Song 2023). However, spatial outliers represent deviations from these patterns, enabling models to adjust predictions based on localized outliers rather than assuming uniform spatial dependence (Lu *et al.* 2009, Zhang *et al.* 2024b). This is particularly important for capturing extreme events, such as variations in agricultural production caused by climate outliers (Anselin 2019). Additionally, spatial outliers reflect variations in explanatory variables across regions, such as differences in soil properties, vegetation indices, or climatic factors, allowing the model to correct better and optimize predictions (Cai *et al.* 2024). By incorporating this additional layer of spatial information, the SDO model improves predictive accuracy and reduces uncertainty (Shen *et al.* 2021).

A key methodological aspect of the SDO approach is its dependence on buffer selection and the definition of outliers, which directly impact model performance. To assess the sensitivity of these parameters, we conducted a comprehensive sensitivity

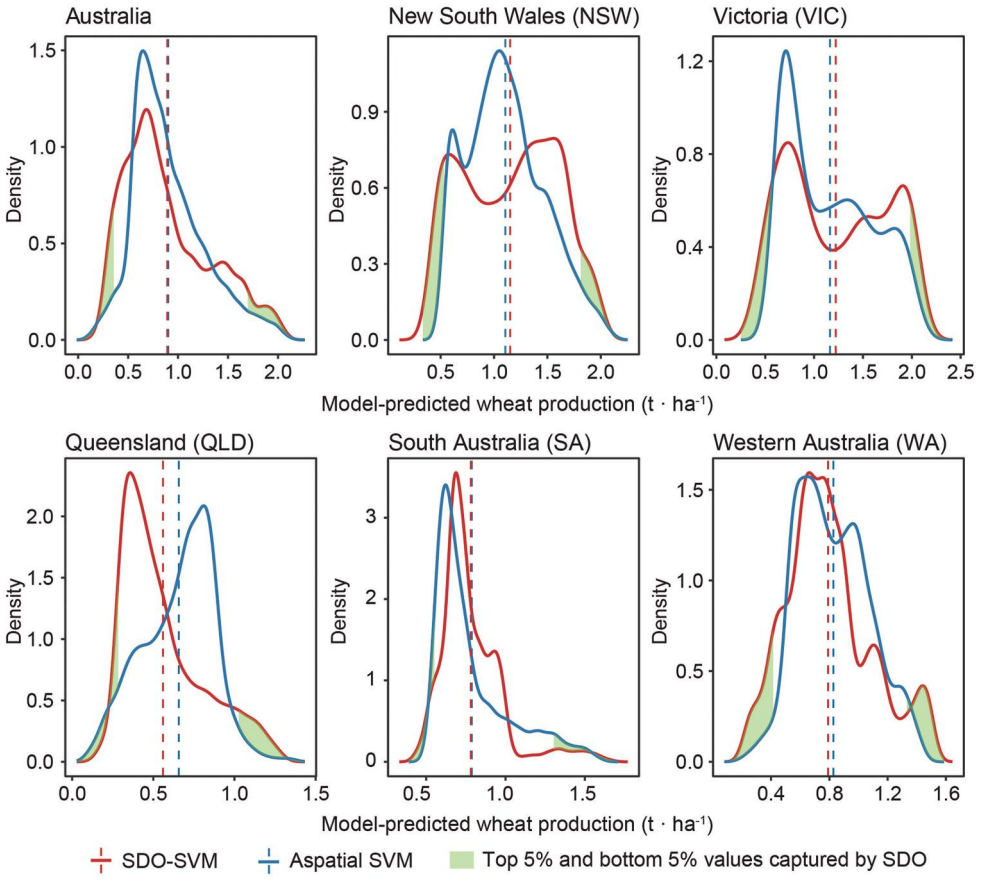

Figure 14. Density distribution map of the average wheat production prediction results of the SDO and aspatial models in Australia and each state.

analysis (Section 4.23) to examine the influence of buffer threshold (m) and interval selection on model accuracy. The results indicate that model performance, measured by cross-validation RMSE, fluctuates by less than 5% across different parameter settings, confirming that the SDO method remains robust under various buffer configurations. Buffer sizes must be carefully selected to balance spatial resolution and computational efficiency. While larger buffers capture broader spatial structures, they may dilute local patterns, whereas smaller buffers emphasize local variations but may fail to account for wider contextual influences. Empirical results suggest that a buffer threshold of 700 km and an interval of 100 km provide an effective balance between multi-scale spatial representation and computational efficiency. Likewise, the selection of the outlier identification threshold plays a crucial role in shaping the SDO variables. In this study, outliers were defined using a mean ± 2 standard deviations ($\overline{x} \pm 2\sigma$) criterion, which ensures a systematic and objective identification of outlier values while preventing excessive sensitivity to minor data fluctuations.

While our choice of five to ten buffers spanning 10–20% of the maximum pairwise distance is guided by empirical practice and supported by recent studies (e.g. buffers of 100–500 m or 250–1,000 m in urban and environmental contexts (Shi *et al.* 2021, Zhang

et al. 2021, Qi *et al.* 2022)), we acknowledge that this range remains heuristic. Optimal buffer sizes may in fact depend on the spatial processes at hand—such as dispersal distances or policy-driven scales—and could be further refined via automated calibration (e.g. cross-validation in GWR/MGWR) or adaptive, location-specific buffers. We therefore recommend that future work explore buffer-size optimization techniques and variable buffer schemes to better tailor the SDO approach to diverse datasets and applications.

The SDO concept extends spatial outlier analysis by integrating multi-scale spatial features, enhancing the model's ability to capture both localized variations and broader spatial correlations. By incorporating multiple buffer ranges, SDO avoids the limitations of single-scale approaches and ensures that spatial outliers are effectively represented. Unlike conventional neighborhood-based models that aggregate or weight all values within a buffer zone, SDO selectively extracts outlier values—outliers—that are more likely to influence the dependent variable. This design helps preserve the signal of local extremes that might otherwise be diluted by full-buffer averaging. While aggregating all spatial context can improve prediction in some cases, it may introduce redundancy, obscure localized deviations, and increase model complexity. Our results demonstrate that the SDO approach, by isolating outlier information across varying distances, offers an efficient and accurate way to represent spatial heterogeneity without unnecessary complexity. Future research may explore hybrid frameworks that combine outlier-based and full-neighborhood strategies to further improve spatial prediction in diverse settings.

Applying the SDO model to wheat production prediction in Australia revealed that several predictor variables exhibit strongly localized outliers, significantly influencing yield variation. Aspatial models fail to capture these spatial outliers, reducing prediction accuracy. By incorporating multi-scale spatial outliers, the SDO model demonstrated notable improvements in predictive performance. For instance, positive NDVI outliers within a 500 km buffer accounted for 32%, and outliers at other buffer distances showed even stronger signals. The SDO model effectively identified these spatial outliers, allowing for a more accurate and context-aware representation of spatial variability. These findings reinforce the effectiveness of SDO in detecting and utilizing spatial outliers, contributing to more robust, precise, and interpretable spatial predictions.

The integration of SDO into spatial modeling provides a powerful mechanism for improving predictive accuracy by leveraging multi-scale spatial variability. The results demonstrate that spatial outliers are critical for refining predictive models, particularly in addressing extreme values, spatial heterogeneity, and nonlinear spatial dependencies. In addition, sensitivity analysis confirms that SDO remains stable across a range of buffer configurations and outlier thresholds, further reinforcing its robustness and adaptability. Overall, the SDO framework offers an important extension to existing spatial prediction methods, enabling a more comprehensive and scalable approach to handling spatial outliers and improving model reliability across diverse geographic applications.

5.2. Limitations and future work

This study still has several limitations. The analysis relates mean wheat production for each LGA located within the Australian wheat belt to climate, soil and environmental variables that are averaged over the entire administrative polygon. This spatial

mismatch raises the ecological fallacy and the modifiable areal unit problem. To lessen these biases, we first restricted every environmental layer to the mapped wheat belt inside each LGA before aggregation so that both response and explanatory data refer to comparable land. Some aggregation error nevertheless remains, so the present results should be viewed as a numerical demonstration of the SDO model rather than an operational decision product. When paddock-scale or fine-grid production data become available, the workflow can move directly to those resolutions and be complemented by high-resolution analyses, multi-scale sensitivity testing, planting-area weighting and fully spatially explicit models that will further mitigate ecological fallacy and modifiable areal unit problem (MAUP) effects.

6. Conclusion

This study proposed the concept of second-dimension outliers (SDO), which identifies positive and negative outliers at various scales within localized buffers based on the spatial location of sample points. These outliers are then used to generate second-dimension outlier variables and develop a second-dimension outlier machine learning model to predict local outliers accurately. The effectiveness and accuracy of the model are validated through case studies involving spatially simulated data and wheat production predictions in Australia. The results demonstrate that the SDO model excels in capturing spatial outliers across multiple scales. By extracting geographic and environmental features from areas beyond the sample points, the model significantly enhances prediction accuracy and avoids blurring outlier information, thereby reducing prediction errors. We recommend that future studies explore more second-dimension spatial characteristics and outliers, such as second-dimension heterogeneity or complexity, to improve current spatial or machine learning models. SDO holds the potential for more accurate and reliable applications across various fields, including agriculture, meteorology, ecology, urban studies, and green space health.

Acknowledgements

We appreciate the Editor-in-Chief, Prof. May Yuan, and the Associate Editor, Prof. Shawn Laffan, as well as the anonymous reviewers, for their insightful comments and suggestions, which have significantly enhanced the quality of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the China Scholarship Council (Grant No. 202206300058).

Notes on contributors

Kai Ren obtained his PhD from Northwest A&F University in China and is currently a postdoctoral researcher at the State Key Laboratory of Climate System Prediction and Risk Management,

Nanjing Normal University, China. Kai was a visiting researcher at Curtin University, Australia. His research interests include geospatial analysis, Crop yield simulation, GIS and agriculture, and theoretical and methodological developments in GIS.

Yongze Song is a Senior Lecturer at Curtin University, Australia. His research interests include spatial statistics, geospatial intelligence, sustainable infrastructure, and sustainable development. He serves as an Associate Editor for the journals GIScience & Remote Sensing, and the International Journal of Applied Earth Observation and Geoinformation.

Qiang Yu is a Professor at the State Key Laboratory of Soil and Water Conservation and Desertification Control, Northwest A&F University, China. His research focuses on agricultural resources and environment, climate change, and agro-ecosystem studies.

ORCID

Kai Ren http://orcid.org/0009-0007-1079-902X
Yongze Song http://orcid.org/0000-0003-3420-9622
Qiang Yu http://orcid.org/0000-0001-6950-1821

Data and codes availability statement

The data and codes that support the findings of this study are available on Figshare at https://doi.org/10.6084/m9.figshare.27245838.

References

Abakar, K.A., and Yu, C., 2014. Performance of SVM based on PUK kernel in comparison to SVM based on RBF kernel in prediction of yarn tenacity. *Indian Journal of Fibre & Textile Research*, 39 (1), 55–59.

Anselin, L., 2019. Quantile local spatial autocorrelation. *Letters in Spatial and Resource Sciences*, 12 (2), 155–166.

Araki, S., *et al.*, 2017. Effect of spatial outliers on the regression modelling of air pollutant concentrations: a case study in japan. *Atmospheric Environment*, 153, 83–93.

Australian Bureau of Statistics. 2021. Value of agricultural commodities produced by local government area - 2020-21. https://public.tableau.com/app/profile/australian.bureau.of.agricultural.and.resource.economics.and.sci/viz/AgCensus_LGA_2020-21/StoryLGA

Baba, A.M., Midi, H., and Abd Rahman, N.H., 2022. Spatial outlier accommodation using a spatial variance shift outlier model. *Mathematics*, 10 (17), 3182.

Berke, O., 2001. Modified median polish kriging and its application to the Wolfcamp–Aquifer data. *Environmetrics*, 12 (8), 731–748.

Cai, J., *et al.*, 2024. Outlier detection in spatial error models using modified thresholding-based iterative procedure for outlier detection approach. *BMC Medical Research Methodology*, 24 (1), 89.

Cherkassky, V., and Ma, Y., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks: The Official Journal of the International Neural Network Society*, 17 (1), 113–126.

Didan, K., 2021. Modis/terra vegetation indices 16-day l3 global 1km sin grid v061. Distributed by NASA EOSDIS Land Processes Distributed Active Archive Center.

Din, S.U., and Yamamoto, K., 2024. Urban spatial dynamics and geo-informatics prediction of Karachi from 1990–2050 using remote sensing and CA-ANN simulation. *Earth Systems and Environment*, 8 (3), 849–868.

Ding, X., *et al.*, 2021. Random radial basis function kernel-based support vector machine. *Journal of the Franklin Institute*, 358 (18), 10121–10140.

Feng, P., *et al.*, 2022. Increasing dominance of Indian Ocean variability impacts Australian wheat yields. *Nature Food*, 3 (10), 862–870.
Georganos, S., and Kalogirou, S., 2022. A forest of forests: a spatially weighted and computationally efficient formulation of geographical random forests. *ISPRS International Journal of Geo-Information*, 11 (9), 471.
Guerschman, J.P., *et al.*, 2022. Estimating actual evapotranspiration at field-to-continent scales by calibrating the CMRSET algorithm with MODIS, VIIRS, Landsat and Sentinel-2 data. *Journal of Hydrology*, 605, 127318.
Harris, P., *et al.*, 2010. The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. *Mathematical Geosciences*, 42 (6), 657–680.
Hersbach, H., *et al.*, 2023. Era5 monthly averaged data on single levels from 1979 to present. *In: Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. Reading, UK: European Centre for Medium-Range Weather Forecasts (ECMWF).
Hu, J., Song, Y., and Zhang, T., 2025. A local indicator of stratified power. *International Journal of Geographical Information Science*, 39 (4), 925–943.
Jia, P., Liu, S., and Yang, S., 2023. Innovations in public health surveillance for emerging infections. *Annual Review of Public Health*, 44 (1), 55–74.
Jiang, Z., 2019. A survey on spatial prediction methods. *IEEE Transactions on Knowledge and Data Engineering*, 31 (9), 1645–1664.
Kim, H.-S., Chung, C.-K., and Kim, H.-K., 2016. Geo-spatial data integration for subsurface stratification of dam site with outlier analyses. *Environmental Earth Sciences*, 75 (2), 1–10.
Kim, Y., *et al.*, 2024. An effective algorithm of outlier correction in space–time radar rainfall data based on the iterative localized analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–16.
Liu, J., *et al.*, 2010. Visual quality recognition of nonwovens using wavelet texture analysis and robust Bayesian neural network. *Textile Research Journal*, 80 (13), 1278–1289.
Liu, X., Chen, F., and Lu, C.-T., 2014. On detecting spatial categorical outliers. *GeoInformatica*, 18 (3), 501–536.
Lu, Q., Chen, F., and Hancock, K., 2009. On path anomaly detection in a large transportation network. *Computers, Environment and Urban Systems*, 33 (6), 448–462.
Lu, W., Atkinson, D.E., and Newlands, N.K., 2017. Enso climate risk: predicting crop yield variability and coherence using cluster-based PCA. *Modeling Earth Systems and Environment*, 3 (4), 1343–1359.
Militino, A., Palacios, M., and Ugarte, M., 2006. Outliers detection in multivariate spatial linear models. *Journal of Statistical Planning and Inference*, 136 (1), 125–146.
Mishra, U., *et al.*, 2017. Spatial representation of organic carbon and active-layer thickness of high latitude soils in cmip5 earth system models. *Geoderma*, 300, 55–63.
Nirel, R., Mugglestone, M.A., and Barnett, V., 1998. Outlier–robust spectral estimation for spatial lattice processes. *Communications in Statistics - Theory and Methods*, 27 (12), 3095–3111.
Overmars, K. d., De Koning, G., and Veldkamp, A., 2003. Spatial autocorrelation in multi-scale land use models. *Ecological Modelling*, 164 (2–3), 257–270.
Qi, M., *et al.*, 2022. National land use regression model for no2 using street view imagery and satellite observations. *Environmental Science & Technology*, 56 (18), 13499–13509.
Rossel, R.V., *et al.*, 2015. The Australian three-dimensional soil grid: Australia's contribution to the globalsoilmap project. *Soil Research*, 53 (8), 845–864.
Sen, Z., 2016. *Spatial modeling principles in earth sciences*. Vol. 10. Cham, Switzerland: Springer.
Shen, X., Bao, W., and Qu, K., 2021. Subspace-based preprocessing module for fast hyperspectral endmember selection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 3386–3402.
Shi, Y., *et al.*, 2018. A graph-based approach for detecting spatial cross-outliers from two types of spatial point events. *Computers, Environment and Urban Systems*, 72, 88–103.
Shi, Y., *et al.*, 2021. A multiscale land use regression approach for estimating intraurban spatial variability of pm2. 5 concentration by integrating multisource datasets. *International Journal of Environmental Research and Public Health*, 19 (1), 321.

Simões, M.V., and Peterson, A.T., 2018. Utility and limitations of climate-matching approaches in detecting different types of spatial errors in biodiversity data. *Insect Conservation and Diversity*, 11 (5), 407–414.
Song, Y., 2022. The second dimension of spatial association. *International Journal of Applied Earth Observation and Geoinformation*, 111, 102834.
Song, Y., 2023. Geographically optimal similarity. *Mathematical Geosciences*, 55 (3), 295–320.
Song, Y., *et al*., 2025. Advancing geospatial methods for addressing global resource and sustainability challenges. *Resources, Conservation and Recycling*, 223, 108517.
Strandberg, J., Sjöstedt de Luna, S., and Mateu, J., 2019. Prediction of spatial functional random processes: comparing functional and spatio-temporal kriging approaches. *Stochastic Environmental Research and Risk Assessment*, 33 (10), 1699–1719.
Sun, X.-L., *et al*., 2019. Performance of median kriging with robust estimators of the variogram in outlier identification and spatial prediction for soil pollution at a field scale. *The Science of the Total Environment*, 666, 902–914.
Taheri Shahraiyni, H., and Sodoudi, S., 2016. Statistical modeling approaches for pm10 prediction in urban areas; a review of 21st-century studies. *Atmosphere*, 7 (2), 15.
Tan, X., *et al*., 2017. Prediction of soil properties by using geographically weighted regression at a regional scale. *Soil Research*, 55 (4), 318–331.
Tang, Y., *et al*., 2024. Robust large-scale collaborative localization based on semantic submaps with extreme outliers. *IEEE/ASME Transactions on Mechatronics*, 29 (4), 2649–2660.
Tian, Y., *et al*., 2022. Forest fire spread monitoring and vegetation dynamics detection based on multi-source remote sensing images. *Remote Sensing*, 14 (18), 4431.
Ulloa-Espindola, R., and Perez-Albert, Y., 2022. Validation of an urban growth prediction model in Quito (Ecuador) built by using weights of evidence and cellular automata. *Eure-Revista Latinoamericana DE Estudios Urbano Regionales 48* (144), 1–27.
Vicente-Serrano, S.M., *et al*., 2023. A global drought monitoring system and dataset based on era5 reanalysis: a focus on crop-growing regions. *Geoscience Data Journal*, 10 (4), 505–518.
Wadoux, A.M.-C., Padarian, J., and Minasny, B., 2019. Multi-source data integration for soil mapping using deep learning. *SOIL*, 5 (1), 107–119.
Wagner, H.H., Chávez-Pesqueira, M., and Forester, B.R., 2017. Spatial detection of outlier loci with Moran eigenvector maps. *Molecular Ecology Resources*, 17 (6), 1122–1135.
Wang, J., Chen, Q., and Chen, Y., 2004. Rbf kernel based support vector machine with universal approximation and its application. In: *International Symposium on Neural Networks*. Berlin, Germany: Springer, 512–517.
Wei, M., Sung, A.H., and Cather, M.E., 2004. *Detecting spatial outliers using bipartite outlier detection methods*. Las Vegas, NV: IKE, 236–244.
Wu, Z., *et al*., 2024. Estimating forest aboveground biomass using a combination of geographical random forest and empirical Bayesian Kriging models. *Remote Sensing*, 16 (11), 1859.
Yang, W., *et al*., 2020. Spatio-temporal patterns of the 2019-NCOV epidemic at the county level in Hubei Province, China. *International Journal of Environmental Research and Public Health*, 17 (7), 2563.
Yin, X., Zhang, W., and Jing, X., 2023. Static-dynamic collaborative graph convolutional network with meta-learning for node-level traffic flow prediction. *Expert Systems with Applications*, 227, 120333.
Yu, H., and Kim, S., 2012. SVM tutorial-classification, regression and ranking. *Handbook of Natural Computing*, 1, 479–506.
Yujun, C., *et al*., 2019. Spatial–temporal traffic outlier detection by coupling road level of service. *IET Intelligent Transport Systems*, 13 (6), 1016–1022.
Zhang, C-T., and Yang, Y., 2019. Improving the spatial prediction of soil Zn by converting outliers into soft data for Bme method. *Stochastic Environmental Research and Risk Assessment*, 33 (3), 855–864.
Zhang, J., *et al*., 2024a. Spatiotemporal meteorological prediction based on fully convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–18.

Zhang, Z., *et al*., 2021. Impacts of land use at multiple buffer scales on seasonal water quality in a reticular river network area. *PloS One*, 16 (1), e0244606.

Zhang, Z., Li, Z., and Song, Y., 2024b. On ignoring the heterogeneity in spatial autocorrelation: consequences and solutions. *International Journal of Geographical Information Science*, 38 (12), 2545–2571.

Zhu, A.-X., *et al*., 2018. Spatial prediction based on third law of geography. *Annals of GIS*, 24 (4), 225–240.

Zhu, H., Hao, H.-K., and Lu, C., 2024. Enhanced support vector machine-based moving regression strategy for response prediction and reliability estimation of complex structure. *Aerospace Science and Technology*, 155, 109634.