Ordinary Kriging and k-Nearest Neighbors Prediction of RGB Pixel Hues

Andy Banks
University of Kansas
GEOL 791 Final Project

Abstract:

This work provides a brief mathematical overview and illustration of the competency of Ordinary Kriging and k-Nearest Neighbors averaging for spatial estimation of pixel hues in an RGB image.

[1] INTRODUCTION

In many fields of physical science, technological and logistical problems often inhibit collection of large, uniformly sampled spatiotemporal datasets. Consequently, it is often desired to estimate the values and associated uncertainty of a spatial variable at unobserved locations.

One approach is the construction of spatial linear models, which seek to provide a description of continuous or categorical variables at unknown locations in space. In the context of geostatistical/spatial models, sampled data is assumed to be the result of a random process. This does not mean that the phenomenon (i.e. aquifer, forest or population) results from a random process. Rather, this assumption provides a basis for estimation and associated uncertainty of spatial variables at unobserved locations. Kriging is a type of spatial linear model commonly used in geostatistics, which estimates the value of a function at a given point, by computing a weighted average of observation values in the neighborhood of the estimation point. Such methods (also called Gaussian Process Regression) are fundamentally a form interpolation. Kriging methods differ from traditional interpolation because they implement a spatial prior covariance function to optimize covariance of interpolated values. This gives the best linear unbiased prediction of intermediate (i.e. unknown) values in a field. This is not the case in traditional interpolation, which implements a piecewise polynomial spline to optimize numerical smoothness of interpolated values. Kriging helps to compensate for spatial/temporal clustering of data, treating a cluster of closely spaced points more like a single data point. Additionally, the availability of uncertainty associated of estimated values provides a basis for stochastic simulation (i.e. generation of multiple alternative realizations of a property). As with all interpolation methods, Kriging is limited by computation cost for large datasets, high uncertainty in estimates subject to sparse observation data, and over/under estimation of high/low values respectively. [5]

Another approach, called k-nearest neighbors (k-NN), is a class of non-parametric methods used in regression and classification problems. k-NN averaging methods find observed samples in the neighborhood of an unobserved estimation point, then either attributes (impute) the value of the closest neighbor directly as a prediction (k=1), or computes a weighted average of the k nearest points (k>1). One reason k-NN is popular in many physical sciences, is that when k=1, prediction values (not location) fall within the bounds of reality, in the sense that they can only take the form of already observed values. Conversely, nonparametric methods such as k-NN do not rely on assumptions from a probability distribution, which inhibit quantification of prediction uncertainty and stochastic simulation. [1]

[2] METHODS

DATA In attempt to capture the complexity and variable nature of typical geologic fields, an RGB image was chosen in place of a randomly simulated field.

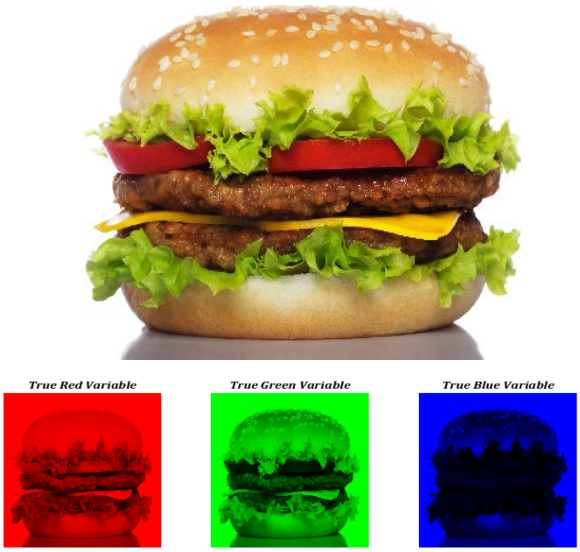


Figure 1: True Field contains Red, Green and Blue pixel hues.

For the purposes of this work, the RGB image will be considered a depiction of some arbitrary physical system, containing three variables; red, green and blue hues [Fig. 1].

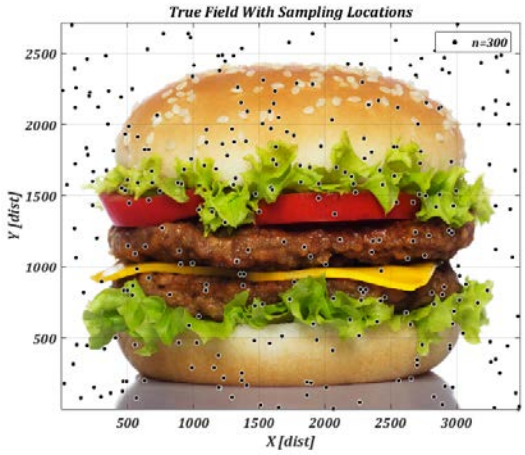


Figure 2: True Field with n=300 pseudo-randomly generated sample locations. 3 observations made at each location; one for each red, green and blue hue respectively

Using Matlab (v2015a), the image was imported as 2710 x 3479 x 3 double precision matrix; each layer containing red, green and blue pixel hues, respectively. Matlab native command randi was used to generate uniformly distributed pseudorandom sampling locations [Fig. 2]. True RGB hue values range between 0 and 1, and sampled values are assumed to have no uncertainty (i.e. void of measurement error). A sample size of 300 was chosen in attempt to capture a physically realistic (0.0003%) amount of information in a true physical field. Given that most physical fields (e.g. earth systems) are continuous in space (i.e. infinitely dense), a question that might be posed; what could be considered a reasonable bound in terms of how densely a continuous field might be measured? 0.0003% is an uneducated guess at what such a number might be, but it's still an interesting idea to think about.

Many statistical methods (including kriging) rely on assumptions of normally distributed sample data. However, in many applications, observed phenomenon are inherently non-normal. Several choices of transformations exist (eg. log, normal-score) that may be useful for normalizing skewed data, though examination these methods are outside the scope of this work.

Histograms in Figure 3 indicate that the sampled RGB data distributions are non-normal, and are generally skewed right (i.e. towards 1). As such, this dataset may represent a less than ideal scenario for estimation methods such as kriging.

VARIOGRAMS Empirical variograms are discrete spatial autocorrelation functions that describe the spatial continuity of an observed phenomenon. Pairs of observation

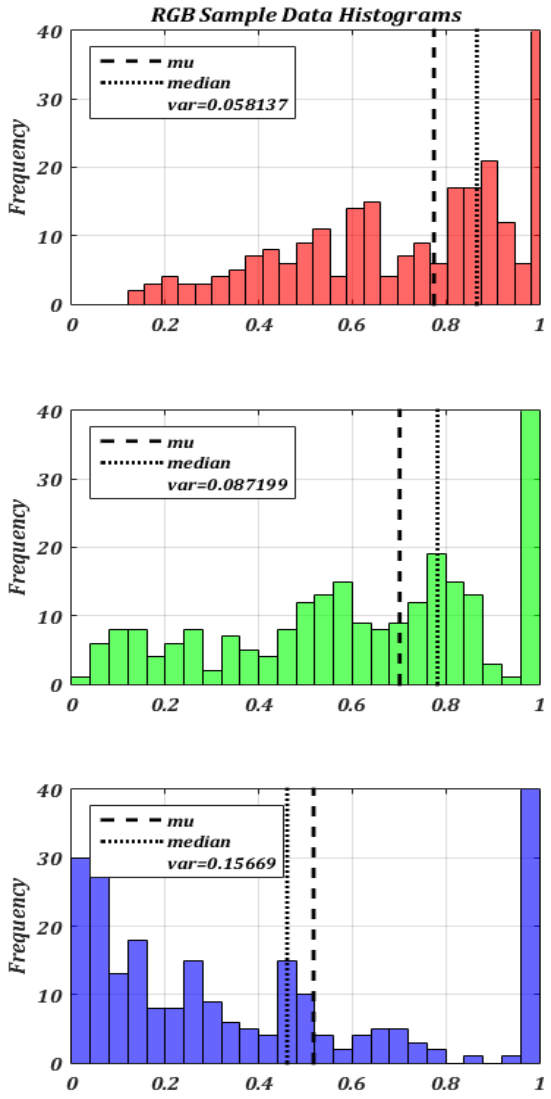


Figure 3: Histograms of RGB hues at sampling loactions in shown in Figure 2.

data are grouped into bins based on the distance between them (called the lag distance or lag vector, h). The variogram is then calculated as the variance of the difference between pairs of observed values in each bin. More explicitly, for observations z_i $\{i = 1, \ldots, n\}$ sampled at locations $u_{a_1}, \ldots, u_{a_n}$, the empirical variogram $\hat{\gamma}(h)$ is defined as

$$\hat{\gamma}(h) \coloneqq \frac{1}{2|N(h)|} \sum_{(i,j)\in N(\boldsymbol{h})} \left|z_i - z_j\right|^2$$

Where $N(h)$ refers to the set (i.e. bin) of observation pairs i, j such that

$$\left|u_{a_i} - u_{a_j}\right| = h$$

and $|N(h)|$ is the numbers of pairs in the set .[3]

The empirical variogram $\hat{\gamma}(h)$ is used to fit an analytical model $\gamma(h)$ of the spatial correlation of the observed phenomenon. Here, $\gamma(h)$ is of spherical form

$$\gamma(h) = \begin{cases} s\left[\frac{3h}{2a} - \frac{\left(h/a\right)^3}{2}\right] & if\ h \le a \\ s & otherwise \end{cases}$$

Where s and a are common geostatistical terms describing parameters of the variogram model: The sill s refers to the limit of the empirical variogram $\hat{\gamma}(h)$ as the lag distance h tends to infinity, and the practical range a to the lag distance h at which the derivative of $\hat{\gamma}(h)$ become negligible (i.e. when the sill is apparent). [5]

KRIGING A Gaussian process is a type of statistical model where observations occur in a continuous domain (e.g. time or space). Every point in the continuous space (i.e. field in this context) of a Gaussian process is associated with a normally distributed random variable. Furthermore, every collection of those random variables has a multivariate normal distribution (i.e. every possible linear combination of the variables is also normally distributed) [2]. The

analytical variogram model $\gamma(h)$ is fundamentally an empirical estimate of the covariance of the Gaussian process governed by the observation data (priors).

Kriging (or Gaussian process regression) is a method of interpolation, modeled by a Gaussian process of prior (i.e. observed/sample data) covariances [3]. Kriging assigns weights $\boldsymbol{\lambda_a}$ for observations $Z^*(\boldsymbol{u_a})$ in accordance with the spatial covariance estimates (i.e. variogram model) $\gamma(h)$; where $Z^*(\boldsymbol{u_a})$ denotes the observed property value z_i , $i = 1, \ldots, n$ sampled at locations $u_{a_1}, \ldots, u_{a_n}$. Assigning weights as such gives the best (minimal variance) linear unbiased estimate of the property Z^*, but only under the assumed spatial covariance model (i.e. Gaussian process). Depending on the stochastic properties of the field, different methods of assigning weights (i.e. methods of kriging) apply. These methods (e.g. simple, ordinary, universal) differ primarily in their treatment of the mean, and the nature of their spatial covariance model [5].

This work will examine Ordinary Kriging, which refers to a spatial prediction of an unknown mean $m(\boldsymbol{u})$, evaluated at each location in $\boldsymbol{u}$. Ordinary Kriging implements a restricted search neighborhood around each estimation location (i.e. using the k nearest neighbors from $\boldsymbol{u_a}$ to compute $m(u_i)$), under the assumption of a constant unknown mean *only* within that search neighborhood**.** The OK estimate is

$$\boldsymbol{Z^*(u)} = \sum_{\boldsymbol{a=1}}^{\boldsymbol{n(u)}} \boldsymbol{\lambda_a} * Z^*(\boldsymbol{u_a})$$

$$\boldsymbol{Z^*(u) = m(u) + \varepsilon}$$

$$\boldsymbol{Z^*(u_a) = m(u_a) + \varepsilon_a}$$

Where $m(\boldsymbol{u})$ and $m(\boldsymbol{u_a})$ are the estimated and observed means of the random process $\boldsymbol{Z^*}$ (e.g. pixel hue), and ε and ε_a are the errors associated with $\boldsymbol{m(u)}$ and $m(\boldsymbol{u_a})$ respectively [5].

K-NEAREST NEIGHBORS k-NN is a non-parametric lazy learning algorithm; meaning that no assumptions are made about the underlying data distribution, and that the function is approximated locally, relying only on training data (e.g. observation data). Consequently, k-NN predictions incorporate no uncertainty information about the priors (i.e. training data) which severely restricts any assessment of prediction uncertainty.
At each evaluation point u_i, $\{i = 1, \ldots, n\}$ in the space of the predictor variables X , ($Z^*(\boldsymbol{u}) \in X$) , k-NN forms the set Y, consisting of the k nearest neighbors of estimation point $Z^*(u_i)$ from the training dataset $Z^*(\boldsymbol{u_a})$. The estimate is then computed as

$$Z^*(u_i) = \frac{1}{k} \sum_{u_{a_j} \in Y} Z^*\left(u_{a_j}\right) \qquad (1)$$

More simply, the estimated property value $Z^*(u_i)$ is a weighted average of its k nearest neighbors in $Z^*(\boldsymbol{u_a})$. If k=1, $Z^*(u_i)$ gets assigned the value of its nearest neighbor in $Z^*(\boldsymbol{u_a})$. If $k=n$ (number of training data), then $Z^*(u_i)$ gets assigned the global average of the training data (i.e. mean). The relationship $\frac{\mathrm{n}}{\mathrm{k}}$ provides a measure of model complexity, with $k=1$ being the most complex and the least complex occurring when $k=n$. Setting $k=1$ is perhaps the *most intuitive* choice of k, as predicted values are certain to be plausible (in terms of value assumed, but not necessarily location) [1]. However, the *best*

choice of k depends upon the data. Generally, larger values of k reduce the effect of noise on the prediction, but make boundaries less distinct between predictions. [5]

EXPERIMENTAL DESIGN Several distinct estimates of the true field are made using both Ordinary Kriging and k-NN methods. The value of k (size of search neighborhood) and the number and spatial distribution of estimation locations in $\boldsymbol{u}$ are varied, but remain consistent between the two methods for each distinct realization (i.e. each distinct of k, $\boldsymbol{u}$ values).
The number of sample locations n (i.e. elements in $\boldsymbol{u_a}$) seems to be the most obvious choice of parameter to vary. However, given that this number comes with physical constraints (i.e. when additional sampling requires manual labor), this value was chosen to remain constant at $n=300$ (equivalent to ~0.0003% of true field).

Accuracy of predictions made by both methods will be compared by sampling the true image at estimation locations, and taking the 2 norm of the differences.

RESULTS

VARIOGRAM MODELS Figure 4 contains example fitted empirical variograms used in Ordinary Kriging estimates.

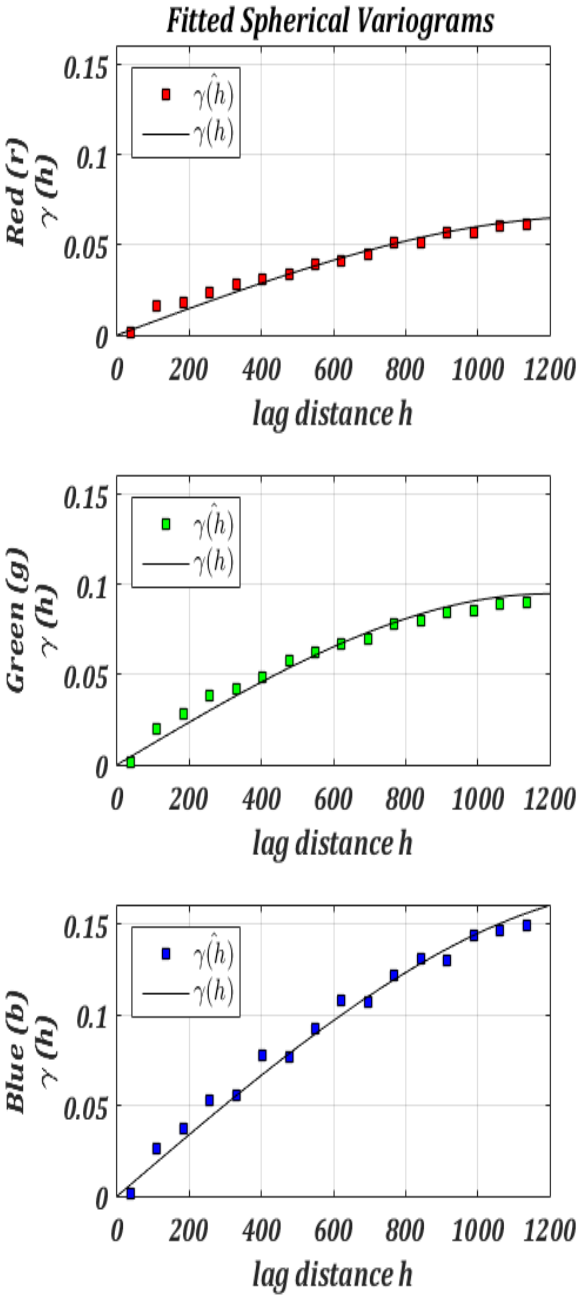


Figure 4: Example empirical variogram and respective spherical models for each variable

CASE A1 *k=1*, using 250000 uniformly spaced estimation points

Figure 5a: Ordinary Kriging (left) and k-NN (right) predictions for each RGB variable, using k=1 at 250000 uniformly distributed sampling locations

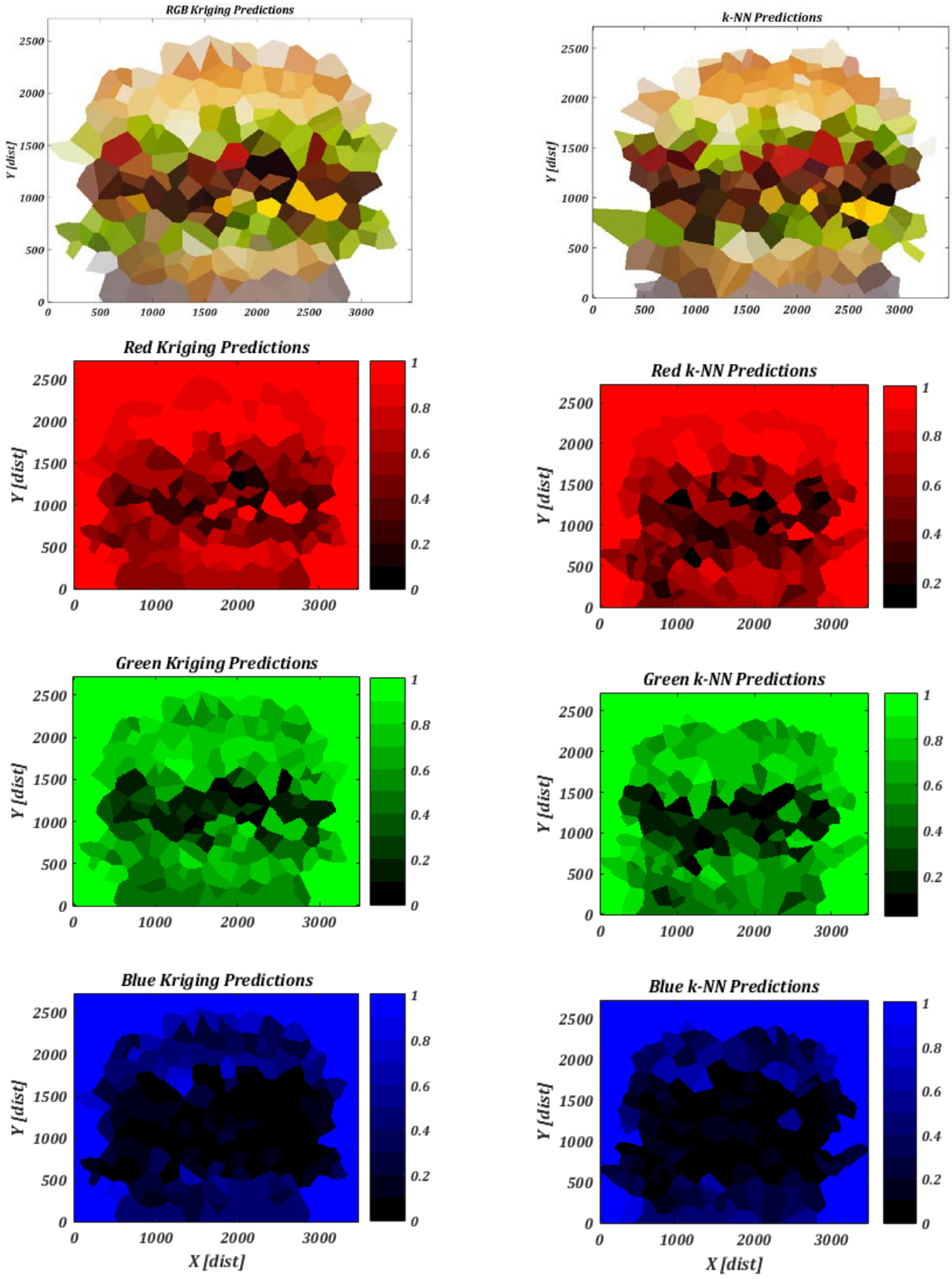
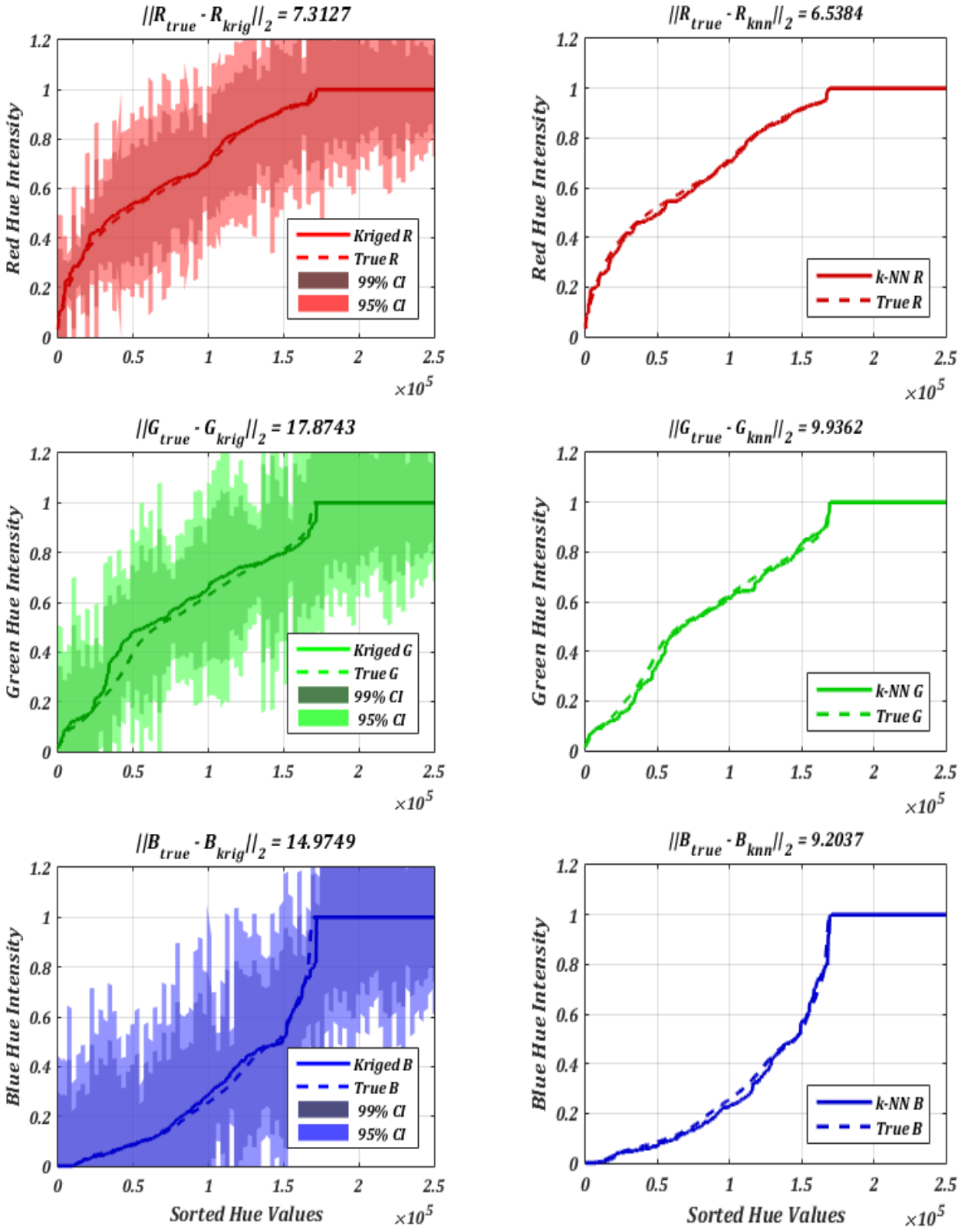


Figure 5b: 2-norm of differences between true field and predictions for Ordinary Kriging and k-NN predictions of each RGB variable, using k=1 at 250000 uniformly distributed sampling locations.



CASE A2 *k=1*, using 2500 uniformly spaced estimation points

Figure 6a: Ordinary Kriging (left) and k-NN (right) predictions for each RGB variable, using k=1 at 2500 uniformly distributed sampling locations

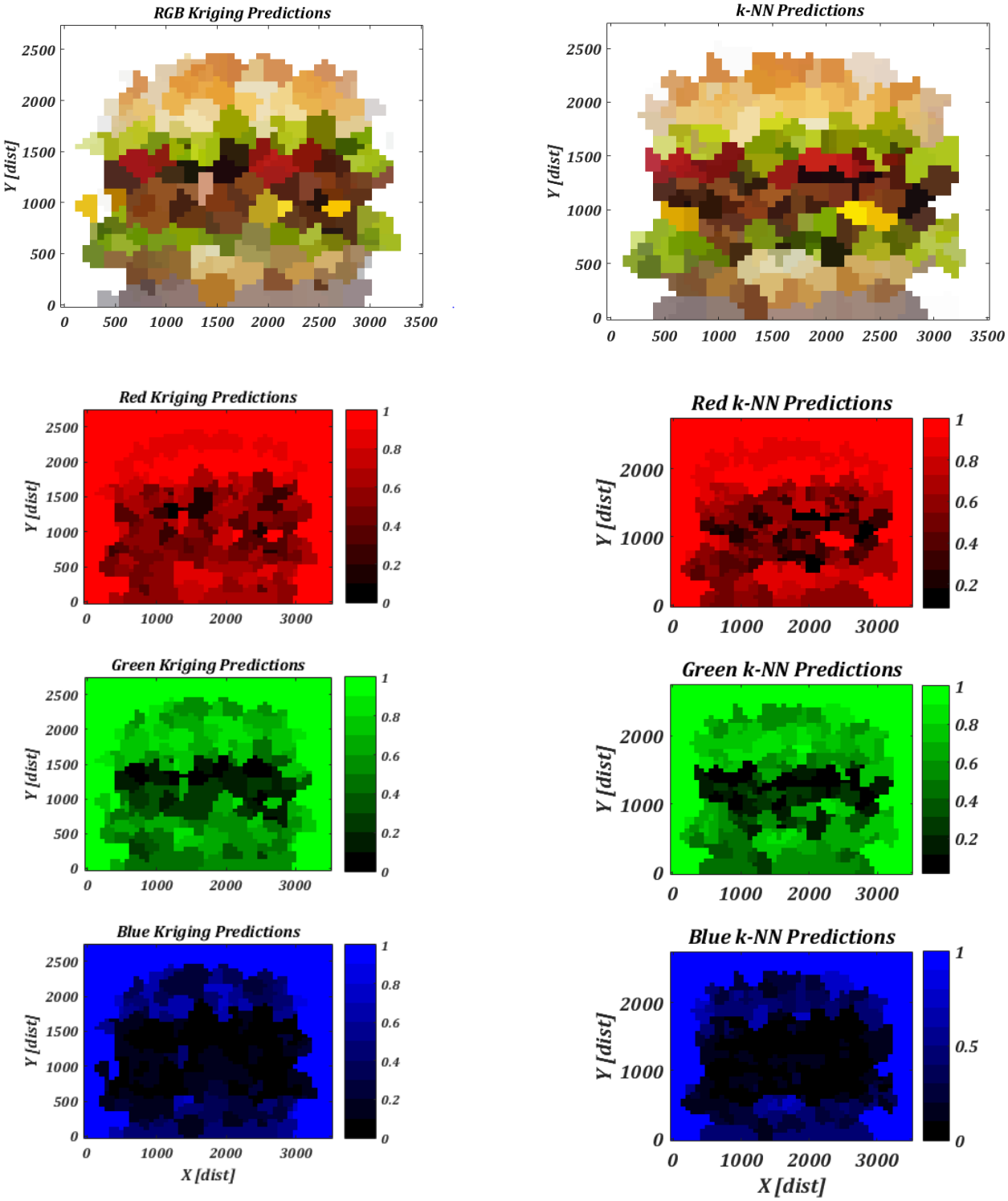
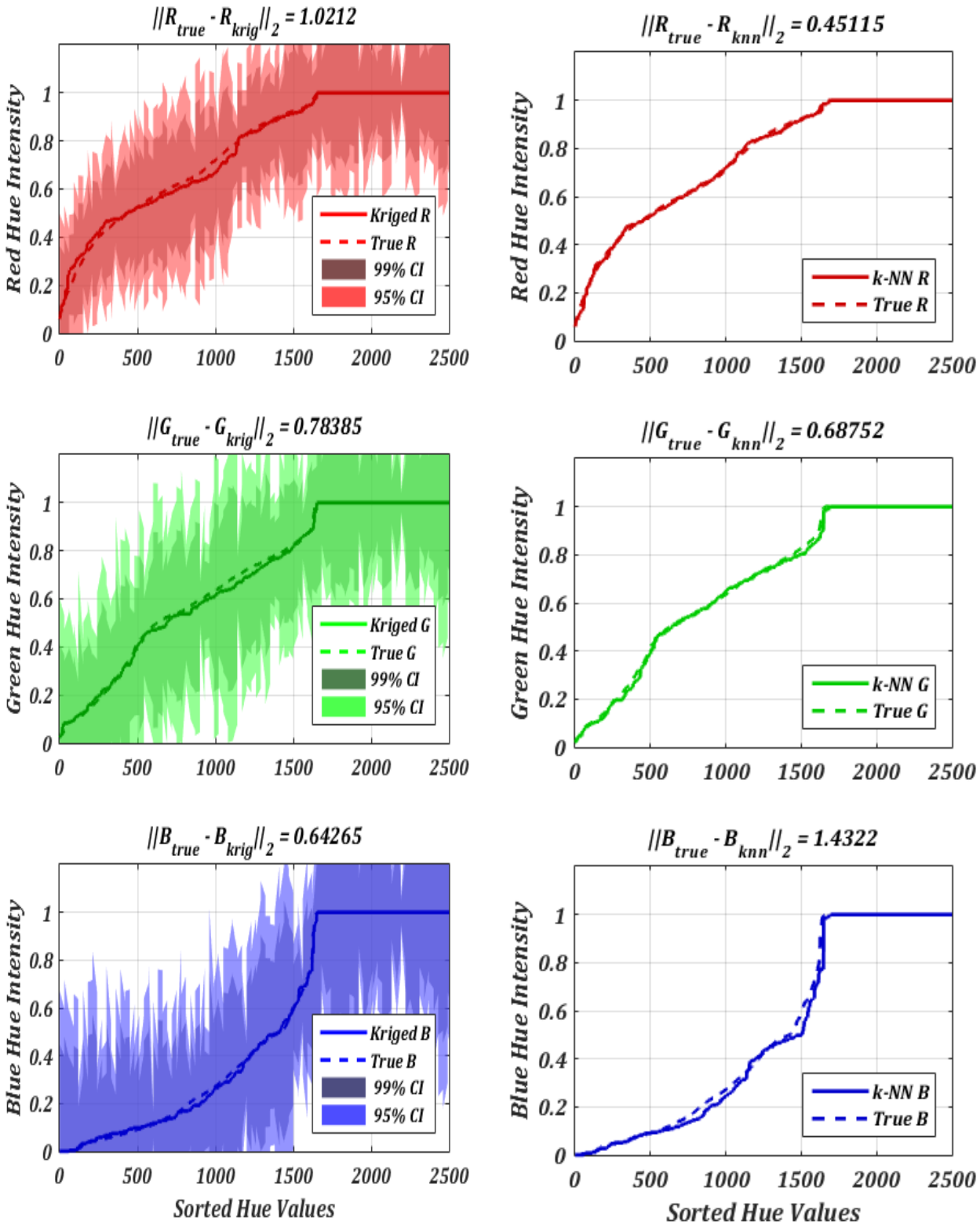


Figure 6b: 2-norm of differences between true field and predictions for Ordinary Kriging and k-NN predictions of each RGB variable, using k=1 at 2500 uniformly distributed sampling locations



CASE B1 $k=3$ using 250000 uniformly spaced estimation points.

Figure 7a: Ordinary Kriging (left) and k-NN (right) predictions for each RGB variable, using k=3 at 250000 uniformly distributed sampling locations

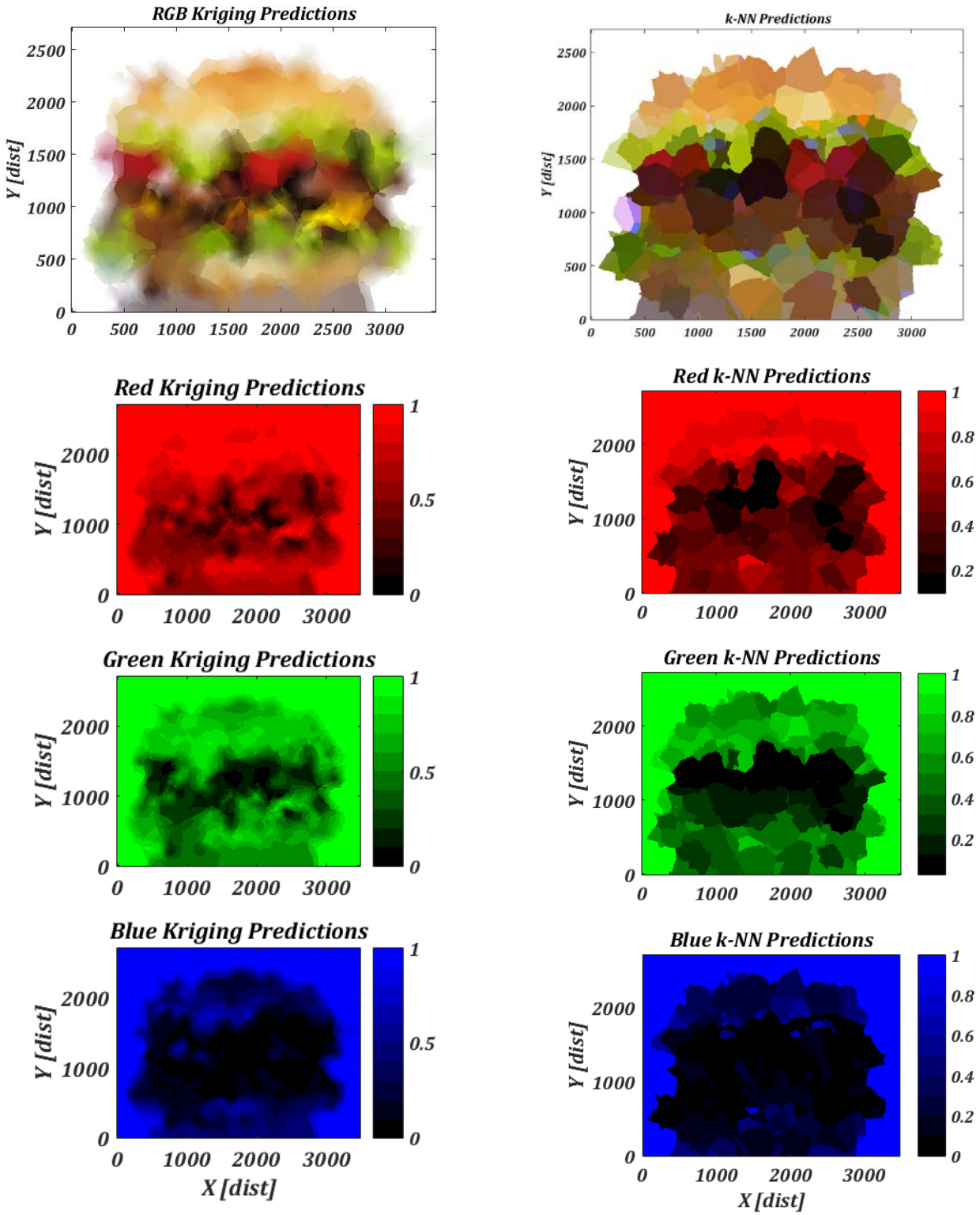
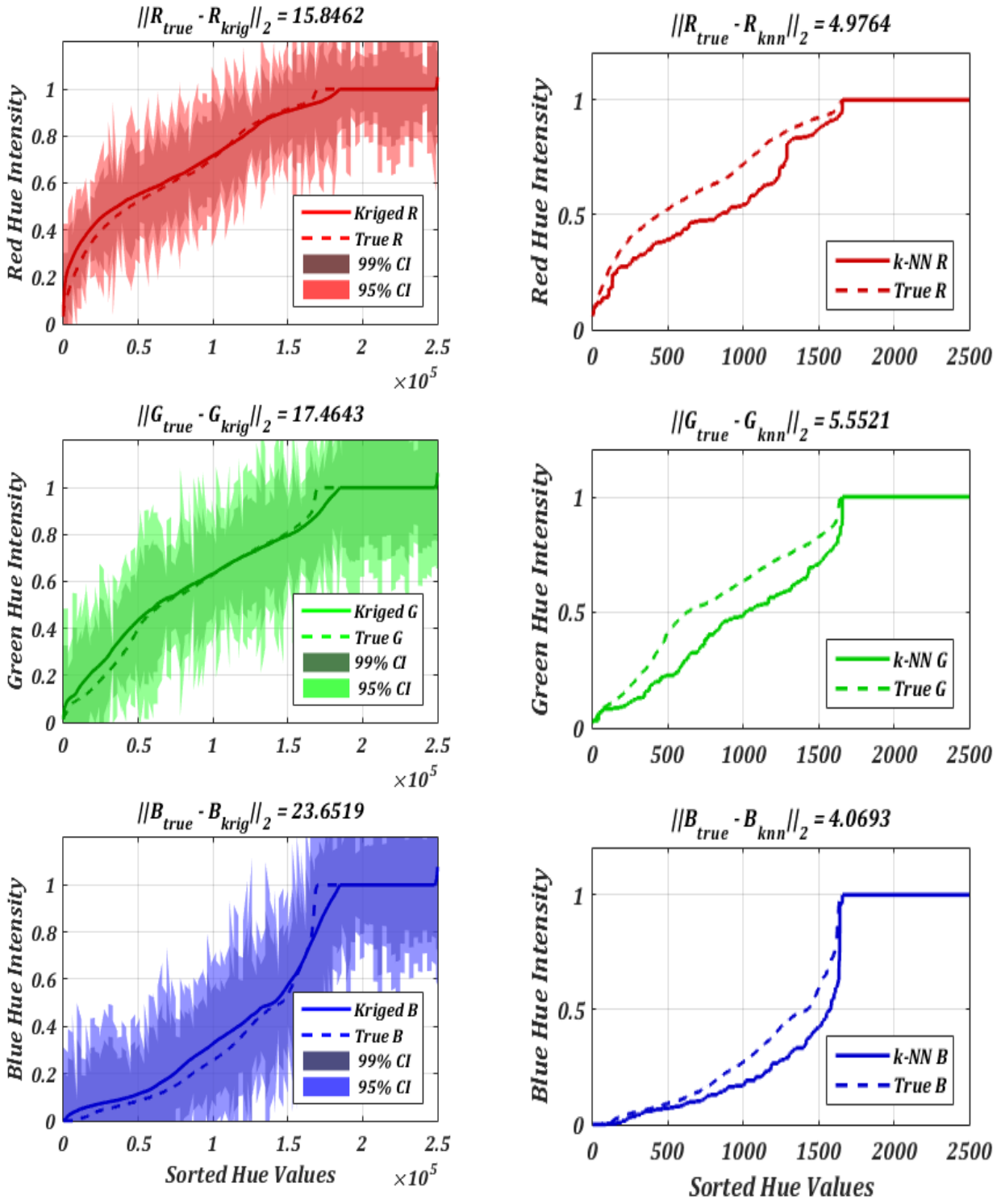


Figure 7b: 2-norm of differences between true field and predictions for Ordinary Kriging and k-NN predictions of each RGB variable, using k=3 at 250000 uniformly distributed sampling locations



CASE B2 k=3 using 2500 uniformly spaced estimation points

Figure 8a: Ordinary Kriging (left) and k-NN (right) predictions for each RGB variable, using k=3 at 2500 uniformly distributed sampling locations

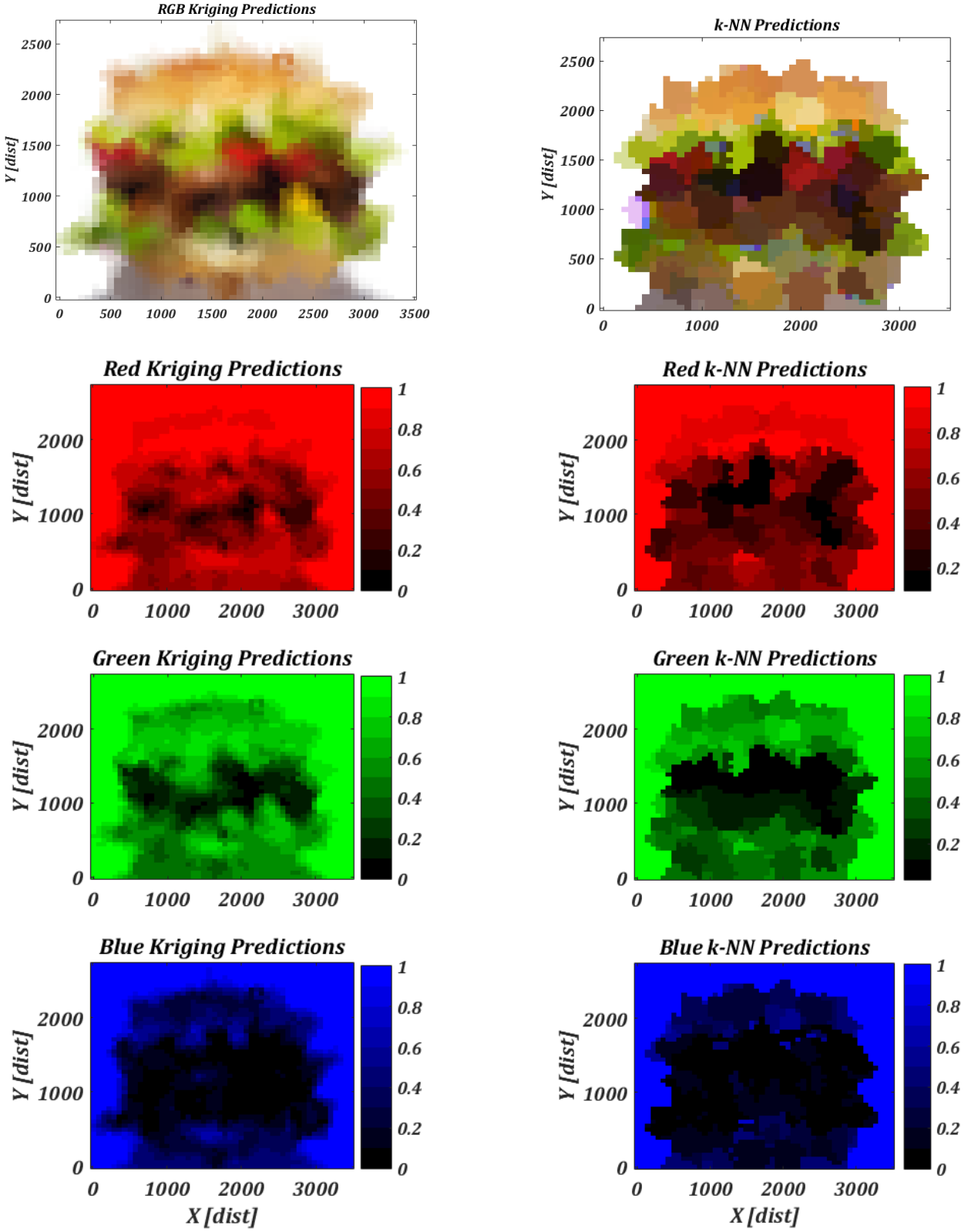
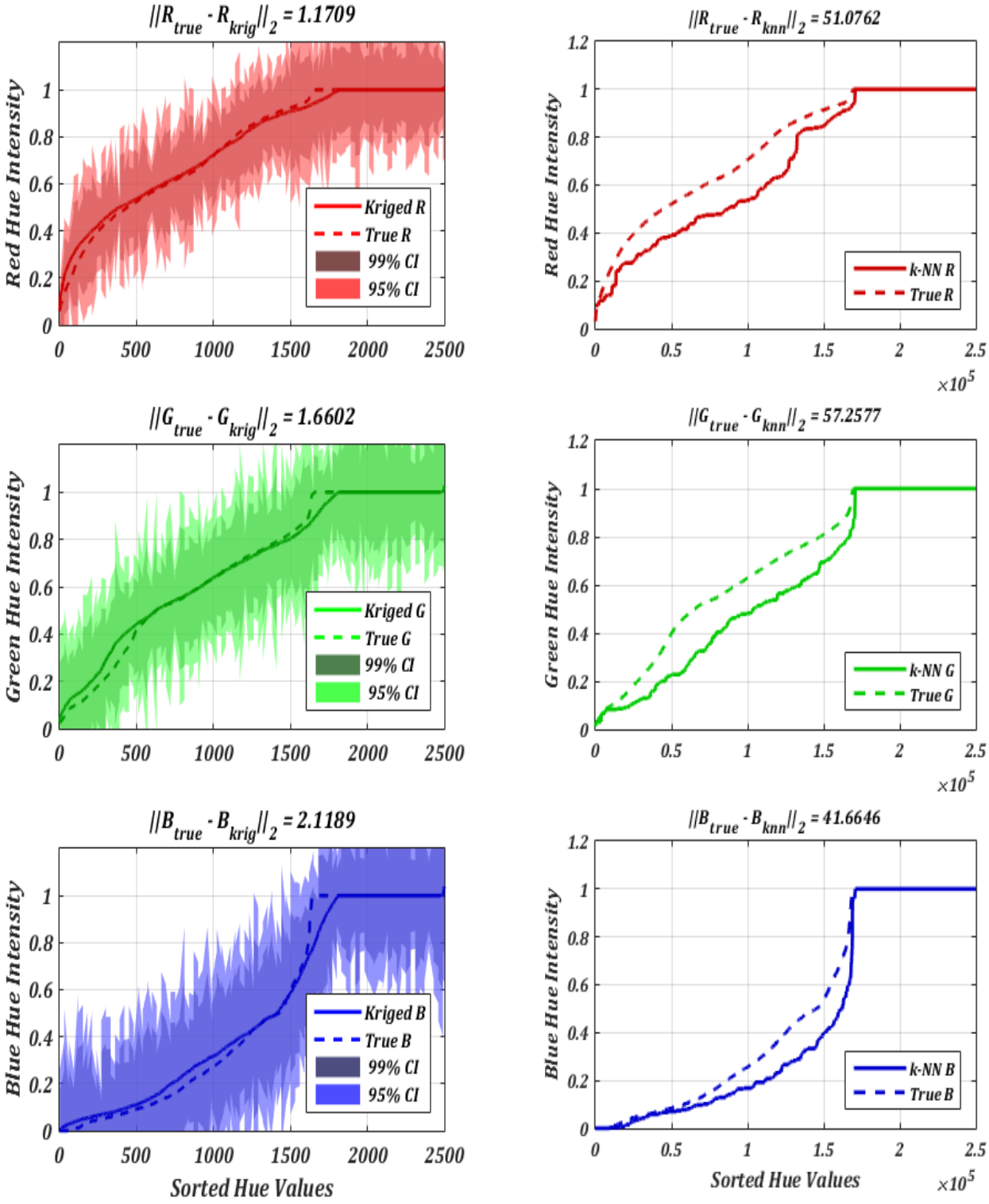


Figure 8b: 2-norm of differences between true field and predictions for Ordinary Kriging and k-NN predictions of each RGB variable, using k=3 at 2500 uniformly distributed sampling locations



CASE C1 *k=50*, using 250000 uniformly spaced estimation points.

Figure 9a: Ordinary Kriging (left) and k-NN (right) predictions for each RGB variable, using k=50 at 250000 uniformly distributed sampling locations

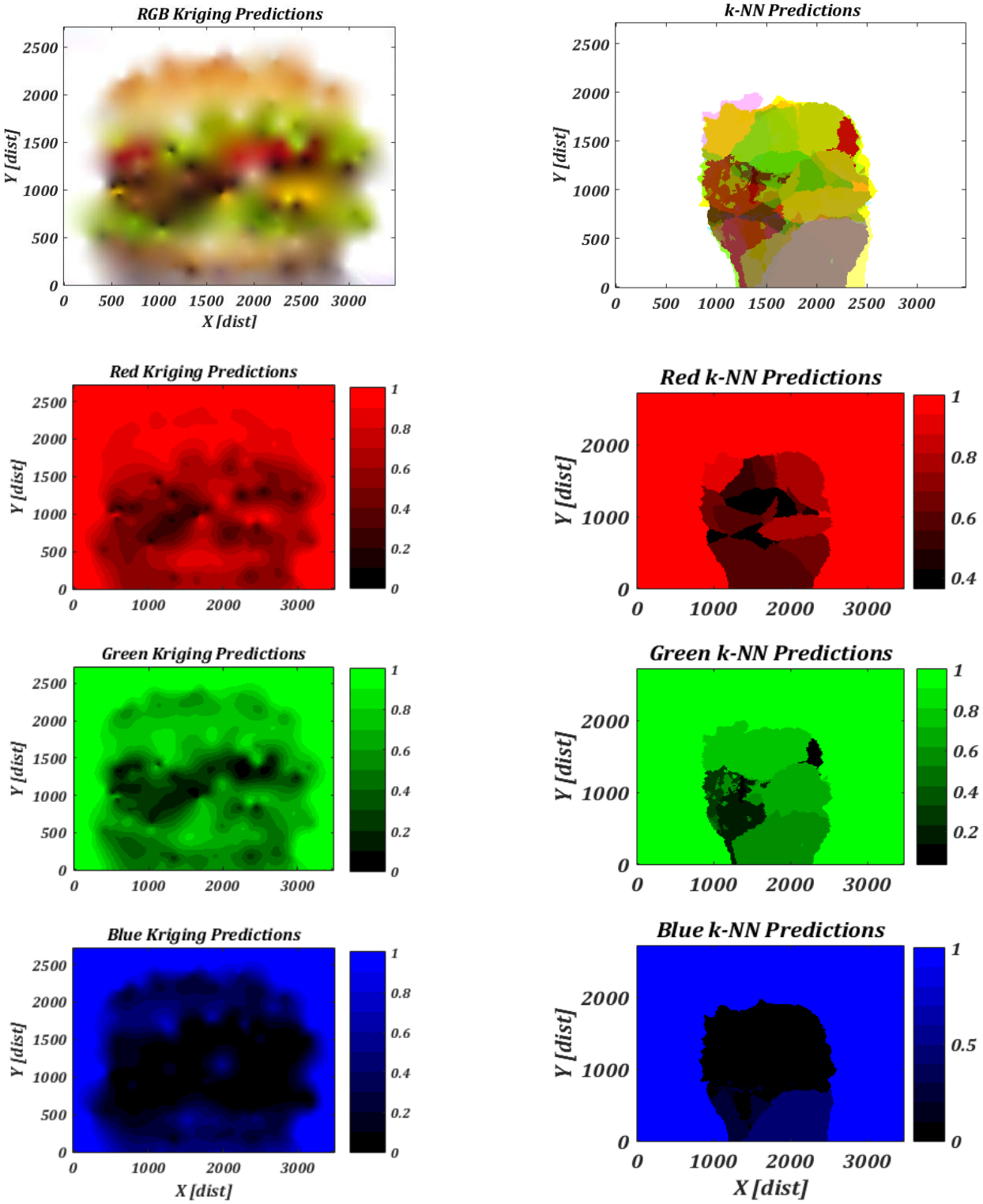
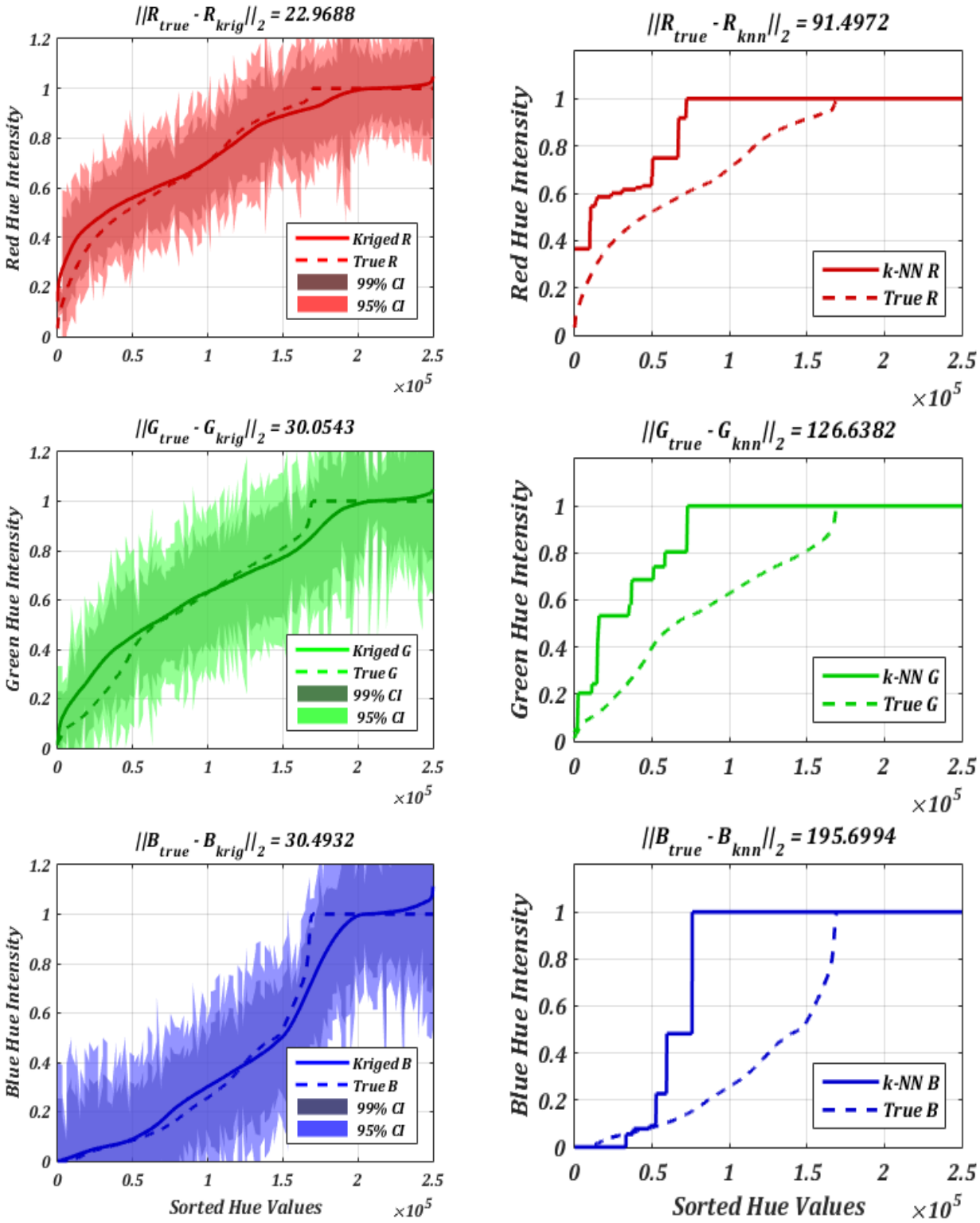


Figure 9b: 2-norm of differences between true field and predictions for Ordinary Kriging and k-NN predictions of each RGB variable, using k=50 at 250000 uniformly distributed sampling locations



CASE C2 k=50 using 2500 uniformly spaced estimation points.

Figure 10a: Ordinary Kriging (left) and k-NN (right) predictions for each RGB variable, using k=50 at 2500 uniformly distributed sampling locations

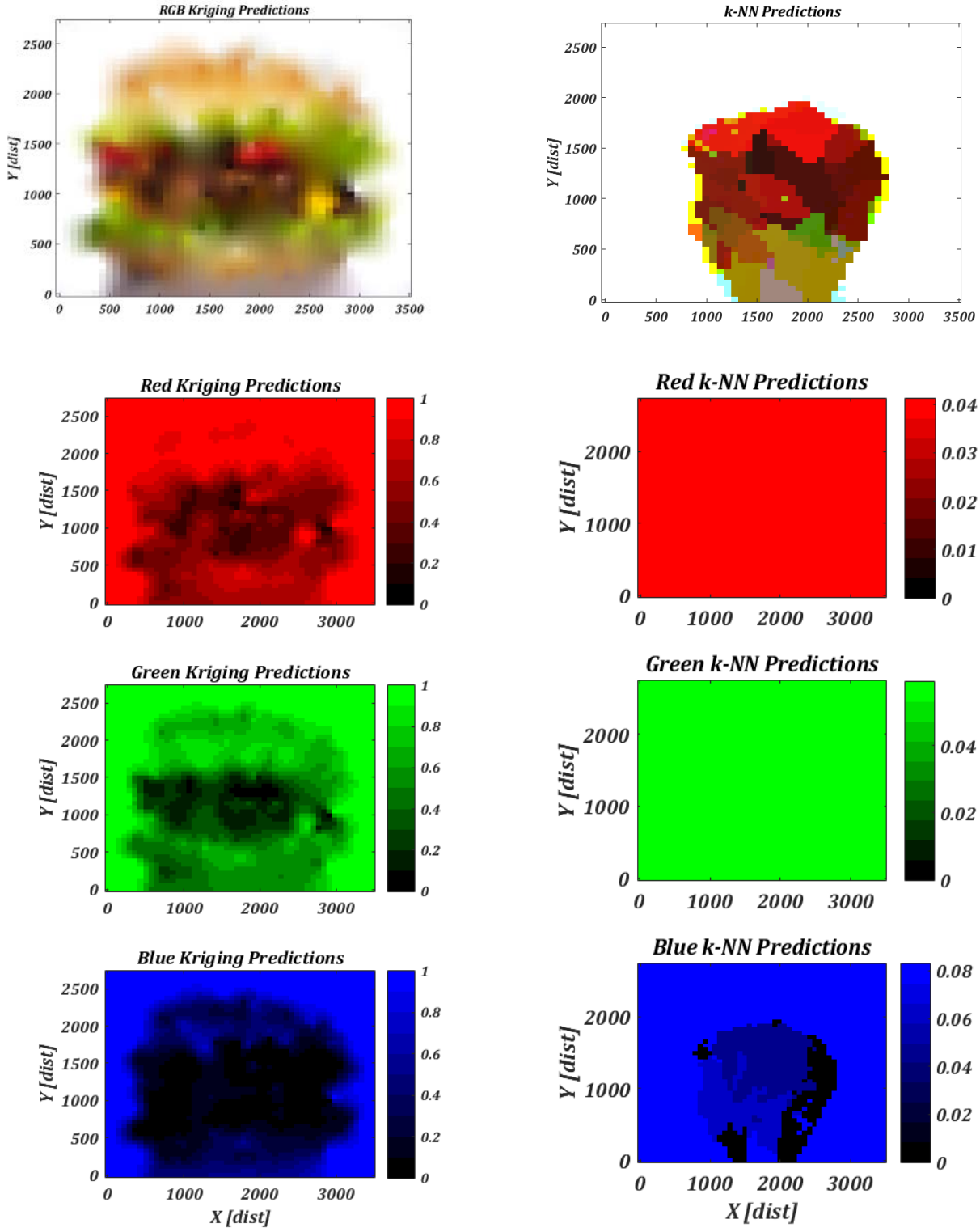
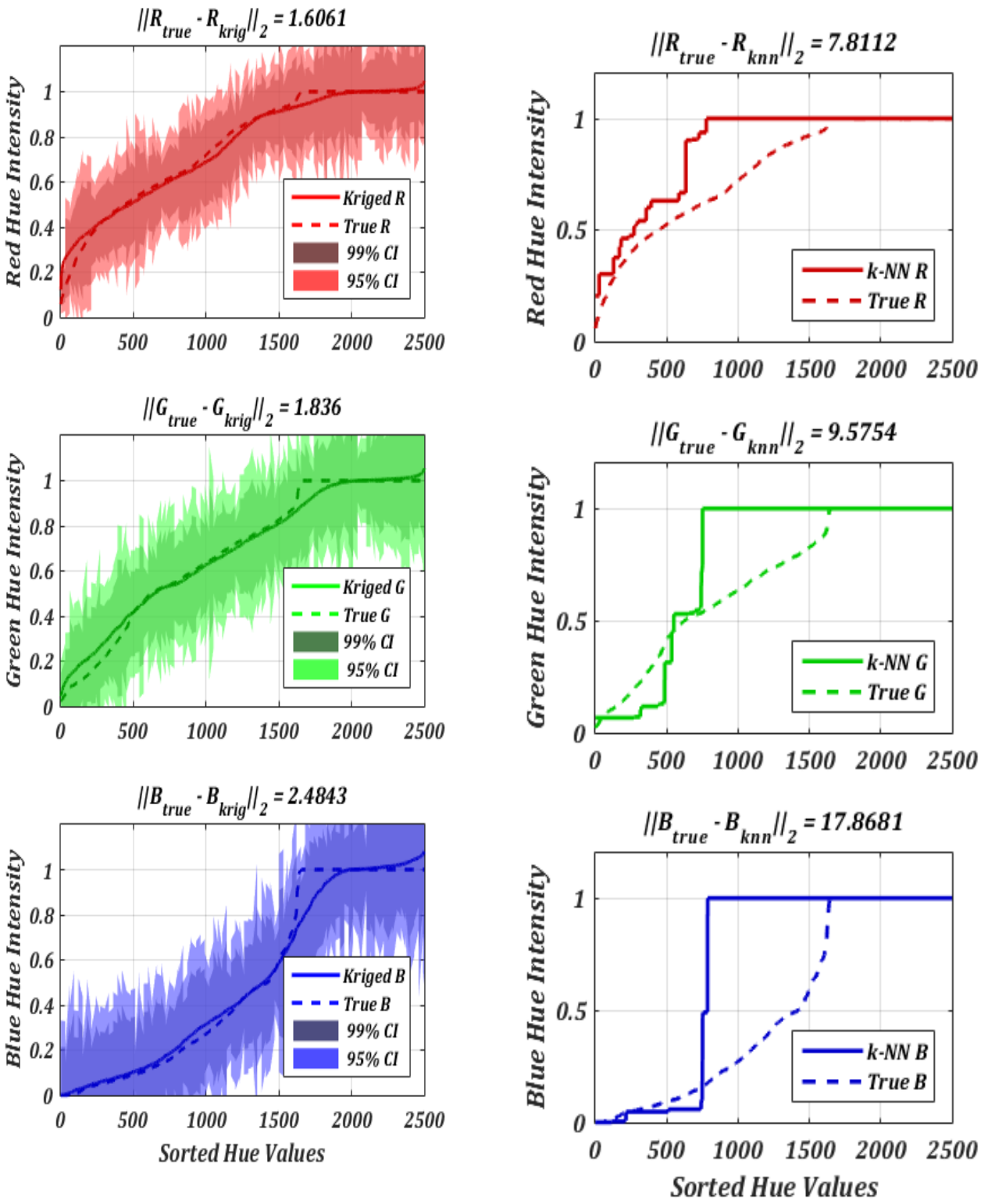


Figure 10b: 2-norm of differences between true field and predictions for Ordinary Kriging and k-NN predictions of each RGB variable, using k=50 at 2500 uniformly distributed sampling locations



DISCUSSION

When $k=1$, ordinary kriging predictions result from interpolation over a neighborhood of the single nearest neighbor, while k-NN predictions are assigned exactly the value of their nearest neighbor in the training dataset. As such, the results from both methods are nearly identical. The best overall estimate, determined by the minimum 2-norm difference of the residual, were produced in case A2. It is likely that predictions for both methods were better in A2 than A1, due the relative difference in sample size to the number of estimation points (i.e. overfitting in case A1).

For $k>1$, k-NN prediction accuracy decreases significantly. Non-plausible values (e.g. purple hues in Figs 7a and 8a) are introduced as a result of the averaging seen in equation 1 (page 5). As k gets very large, k-NN predictions uniformly approach the global mean of the dataset. Changing the number of estimation points has little impact on prediction accuracy. For large *k*, ordinary kriging performs significantly better than k-NN, but suffers in accuracy relative to ordinary kriging with small k.

In general, using a small number of sampling points produced uniformly better results for both methods (A2 B2 C2).

Although these are interesting results, they are flawed. The sampling locations did not remain consistent between all cases, meaning that prediction differences between methods in all cases (A B C) may result from sampling location, rather than the methods themselves. Computation cost prevented this problem from being resolved.

REFERENCES

[1]
Ver Hoef JM, Temesgen H (2013) A Comparison of the Spatial Linear Model to Nearest Neighbor (k-NN) Methods for Forestry Applications.

[2]
Seeger, Matthias (2004). "Gaussian Processes for Machine Learning". International Journal of Neural Systems.

[3]
Cressie A. C. Noel (1993). “Statistics for Spatial Data”. Wiley New York. Chapters 3.2, 6.3

[4]
Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999). When is “Nearest Neighbor” Meaningful?

[5]
Bohling Geoff (2017). Lecture Notes. Geology 791. University of Kansas.