Deriving enhanced geographical representations via similarity-based spectral analysis: predicting colorectal cancer survival curves in Iowa

Michael T. Lash*

Department of Computer Science,
University of Iowa,
Iowa City, IA, USA
Email: michael-lash@uiowa.edu
*Corresponding author

Min Zhang

Interdisciplinary Graduate Program in Informatics,
University of Iowa,
Iowa City, IA, USA
Email: min-zhang@uiowa.edu

Xun Zhou and W. Nick Street

Management Sciences Department,
University of Iowa,
Iowa City, IA, USA
Email: xun-zhou@uiowa.edu
Email: nick-street@uiowa.edu

Charles F. Lynch

Department of Epidemiology,
University of Iowa,
Iowa City, IA, USA
Email: charles-lynch@uiowa.edu

Abstract: Neural networks are capable of learning rich, nonlinear feature representations shown to be beneficial in many predictive tasks. In this work, we use such models to explore different geographical feature representations in the context of predicting colorectal cancer survival curves for patients in the state of Iowa, spanning the years 1989 to 2013. Specifically, we compare model performance using *area between the curves* (ABC) to assess (a) whether survival curves can be reasonably predicted for colorectal cancer patients in the state of Iowa, (b) whether geographical features improve predictive performance, (c) whether a simple binary representation, or a richer, spectral analysis-elicited representation perform better, and (d) whether spectral analysis-based representations can be improved upon by leveraging geographically-descriptive features. In exploring (d), we devise a similarity-based spectral analysis procedure, which allows for the combination of geographically relational and geographically descriptive features. Our findings suggest that survival curves can be reasonably estimated on average, with

predictive performance deviating at the five-year survival mark among all models. We also find that geographical features improve predictive performance, and that better performance is obtained using richer, spectral analysis-elicited features. Furthermore, we find that similarity-based spectral analysis-elicited representations improve upon the original spectral analysis results by approximately 40%.

Keywords: geographical representations; spectral analysis; deep learning; spectral clustering; neural networks; colorectal cancer; survival curve.

Reference to this paper should be made as follows: Lash, M.T., Zhang, M., Zhou, X., Street, W.N. and Lynch, C.F. (2018) 'Deriving enhanced geographical representations via similarity-based spectral analysis: predicting colorectal cancer survival curves in Iowa', *Int. J. Data Mining and Bioinformatics*, Vol. 21, No. 3, pp.183–211.

Biographical notes: Michael T. Lash is currently a PhD candidate in the Department of Computer Science at the University of Iowa, advised by W. Nick Street and Alberto M. Segre. His research interests are broadly in the areas of machine learning and data mining methodology, with specific interests lying in causal learning, learning from graphs, geographical and spatial data mining, among others. He has published in top data mining conferences, such as SDM, top application venues, such as ICHI and BIBM, and top business journals, such as *JMIS*. He has also been the recipient of numerous awards, including the University of Iowa Graduate College Post-Comprehensive Research Fellowship, an NSF GRFP Honourable Mention, and a variety of student travel awards.

Min Zhang is currently a graduate student in the Interdisciplinary Graduate Program in Informatics at the University of Iowa. He received the bachelor degree in Business Analytics and Information Systems from the University of Iowa, in 2017. His research interests include business data analytics, big data management, and Geographic Information Systems (GIS).

Xun Zhou is currently an Assistant Professor in the Department of Management Sciences at the University of Iowa. He received a PhD degree in Computer Science from the University of Minnesota, Twin Cities in 2014. His research interests include big data management and analytics, spatial and spatio-temporal data mining, and Geographic Information Systems (GIS). His works have been published in top conferences and journals such as *ACM SIGKDD*, *IEEE ICDM*, *ACM SIGSPATIAL*, and *IEEE TKDE*. He has received three best paper awards. He was also a co-editor-in-chief of the Springer Encyclopedia of GIS, 2nd edition.

W. Nick Street is the Henry B. Tippie Research Professor and Departmental Executive Officer in the Management Sciences Department at the University of Iowa, with joint appointments in Computer Science, Nursing, and Informatics. He is also the director of the interdisciplinary graduate program in Health Informatics. His research interests are in algorithmic approaches to machine learning and data mining, particularly the use of mathematical optimisation in inductive learning techniques. His recent work has focused on counterfactual reasoning, ensemble construction methods, and personalised healthcare decision making. He has published over 110 journal, conference and workshop papers, and has received an NSF CAREER award and an NIH INRSA postdoctoral fellowship.

Charles F. Lynch is a Professor with a joint appointment in the Department of Epidemiology in the College of Public Health and in the Department of

Pathology in the College of Medicine at The University of Iowa. He has been a faculty member at The University of Iowa since he completed his pathology training in 1986. Since 1990, he has been Principal Investigator of the State Health Registry of Iowa, Iowa's state-wide cancer surveillance program. His primary research interests include cancer surveillance, cancer epidemiology, and environmental epidemiology.

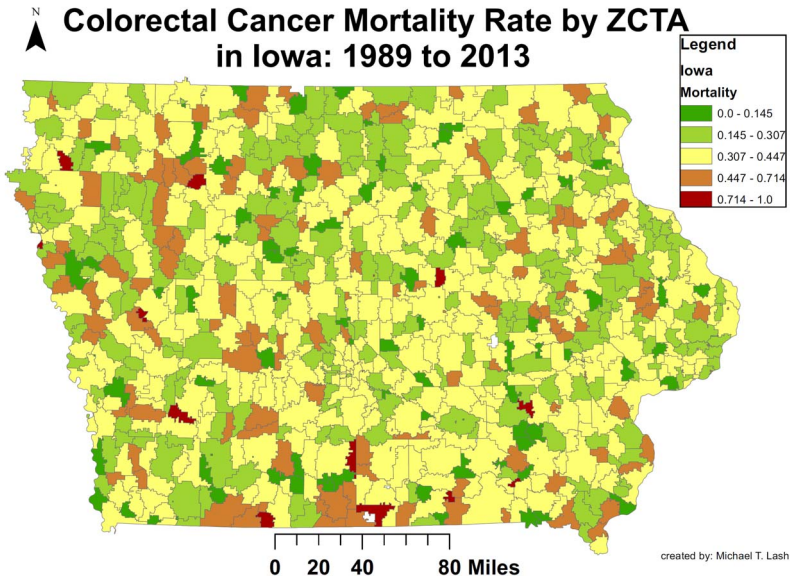
This paper is a revised and expanded version of the paper entitled 'Learning Rich Geographical Representations: Predicting Colorectal Cancer Survival in the State of Iowa' presented at the '2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'17)', Kansas City, MO, 13–16 November 2017.

1 Introduction

As machine learning has become more prevalent, powerful new technologies such as deep learning, which are capable of learning rich, non-linear representation, have also risen to the forefront of the field. The domains of public health and medicine have particularly benefited from these innovations; in this work we examine and propose deep learning methodologies applied to these areas. The focus of this work, therefore, is to explore how different geographical representations, learned through deep learning technologies, can improve survival curve predictions for colorectal cancer patients in the state of Iowa.

Figure 1 demonstrates the urgency of the problem we are addressing, showing colorectal cancer (CRC) mortality rates for patients in Iowa spanning the years 1989 to 2013; these are expressed in terms of a zip code tabulation area (ZCTA) level of geography.

Figure 1 Colorectal cancer mortality rate by ZCTA in the state of Iowa for the years 1989 to 2013



In Figure 1 we first observe that numerous ZCTAs have CRC mortality rates that are at or above 30%, indicating the particularly nefarious nature of this disease, and highlighting the need for accurate survival outlook predictions at the time of diagnosis to better inform treatment decisions (Zhang et al., 2015). Furthermore, Figure 1 demonstrates the geographical diversity in which CRC mortality rates are manifested: locale seems to be related to survival outlook.

The relationship between geography and survival outlook isn't unforeseen, unfortunately. Physical locale manifests pertinent health-based factors, such as access to healthcare, environmental factors, such as ground contaminants, among others, all of which may affect disease manifestation and survival outlook (Wan et al., 2013).

Provided the spatially heterogeneous manifestation of colorectal cancer mortality, a major challenge is to build spatially responsive models that can aid in accurate prediction of individual-specific colorectal cancer survival curves. For instance, rural areas may have a different variety of factors affecting colorectal cancer disease manifestation and survival than sprawling metropolitan cities. Therefore, we define and examine three geographical deep learning representation methods in this work: a *simple binary representation* (SBR), *rich representation–spectral analysis* (RR-SA), and rich representation – similarity-based spectral analysis (RR-SSA); we additionally craft two sub-representation methods that are utilised in RR-SSA.

The contributions of this work, which expand upon the results obtained in Lash et al. (2017b), are enumerated as follows:

1 We investigate a rich representation of spatial features through spectral analysis (RR-SA) of the underlying geographical relationship graph of the ZCTAs to address the spatial heterogeneity challenge.

2 Modifying our RR-SA representation procedure, we explore the use of geographically descriptive features, paired with the underlying adjacency graph, to further address the spatially heterogeneous nature of the problem.

3 We determine whether the simple binary representation (SBR) or richer, spectral analysis representation (RR-SA), or similarity-based spectral analysis representation (RR-SSA) leads to more accurate survival curve predictions.

4 We determine whether RR-SA or RR-SSA representations lead to more accurate survival curve predictions and determine which sub-representation procedure – binary (bin) or full – produces more accurate survival curve predictions.

This works continues with a disclosure of the problem setting, followed by relation of our three methods of geographic representation and two sub-representation methods; we also present a graphic depicting the deep architecture of each method (Section 2). In Section 3, we describe our colorectal cancer patient dataset, containing 46,000 individuals residing in Iowa at the time of their diagnosis; the dataset spans the years 1989 to 2013. Furthermore, we relate our geographical feature dataset, along with our experiments. In Section 4 we disclose works related to ours prior to concluding the paper in Section 5.

2 Learning geographical representations for survival curve prediction

Prior to disclosing our methodology, we relate some preliminary notation, subsequently discussing and mathematically formulating the problem setting. Following this disclosure we reformulate the problem as one of Kaplan-Meier survival curve prediction before introducing and elaborating on our three methods of geographical representation learning and two sub-representations.

2.1 Preliminaries

Define $\{(\mathbf{x}^{(i)}, e^{(i)}, t^{(i)})\}_{i=1}^{n}$ to be a dataset of n instances, where feature vector $\mathbf{x}^{(i)} \in \mathbb{R}^m$, event label $e^{(i)} \in \{0,1\}$, and time of event occurrence $t^{(i)} \in \{0,1,\ldots,T\}$; where $t^{(i)}$ represents a discrete time at which an event $e^{(i)}$ has occurred (i.e., $e^{(i)} = 1$) *or* the last discrete time instance i is observed, while an event has not occurred (i.e., $e^{(i)} = 0$). In this latter case ($e^{(i)} = 0$), when $t^{(i)} = T$ we know the event never occurs to the instance during the study period (spanning T discrete time periods). If, on the other hand, $t^{(i)} < T$ then we only know that the instance did not experience the event up to $t^{(i)}$, but don't know what happened during the $T - t^{(i)}$ remaining time. Representation of event-time data described as such are called *censored data* and, even more specifically, *right-censored* data. A censoring of instance i occurs when $e^{(i)} = 0$ and $t^{(i)} < T$. We elaborate on the handling of these censored data in a proceeding subsection.

To be more concrete, $t \in \{1,\ldots,T\}$ may represent (as is the case in our experiments) six-month patient follow-up periods, with $t = 0$ designating the entrance of a patient to the study. Study entrance occurs when a diagnosis of colorectal cancer is rendered. When an instance (i.e., patient) i dies from colorectal cancer – $e^{(i)} = 1$ – then $t^{(i)}$ designates a time in which this event occurred. Alternately, a patient may pass away from complications not related to their colorectal cancer disease, or may move elsewhere, switch doctors, or for some other reason become untrackable prior to the conclusion of the study period, then $t^{(i)} < T$ and $e^{(i)} = 0$, indicating a censoring.

Patient instance vectors $\mathbf{x}^{(i)}$ represent quantified measurements of pertinent patient-based features. Later in this work, we will make reference to certain feature groups of which these instance vectors are composed. Therefore, we define notation that will conveniently relate to these groups. To such an end, let $\mathbf{z}$ denote the full set of index values that reference the geographical features of $\mathbf{x}^{(i)}$; further, denote $\mathbf{a}$ to be the full set of index values of $\mathbf{x}^{(i)}$ such that $\mathbf{a} = \{1,\ldots,m\}$. We will use these index sets to make direct reference to the feature grouping components of $\mathbf{x}^{(i)}$; for instance, $\mathbf{x}_{\mathbf{z}}^{(i)}$ is the subvector of instance i housing the geographical feature values. Furthermore, using set difference notation, $\mathbf{x}_{\mathbf{a}\setminus\mathbf{z}}^{(i)}$ refers to the subvector of instance i containing feature values that are non-geographical.

For convenience, we provide the notation related in this and subsequent sections in Table 1.

Table 1 Notation used throughout this work

Notation	*Description*
$\mathbf{x}^{(i)} \in \mathbb{R}^m$	Feature vector of instance i.
$e^{(i)} \in \{0,1\}$	Event label of instance i.
$t^{(i)} \in \{1,\ldots,T\}$	Discrete time of $e^{(i)}$.
$\mathbf{y}^{(i)} \in [0,1]^T$	Outcome vector of instance i.
$\hat{\mathbf{y}}^{(i)} \in [0,1]^T$	Predicted outcome vector of instance i.
$\mathbf{z}$	Set of geographical feature index values.
$\mathbf{a}$	Set of all feature index values.
$\mathcal{M}$	A map.
$\Gamma(\cdot)$	Function that determines discrete geographic entity membership.
$P(\cdot)$	Calculation of a probability.
$\mathrm{g}: \mathbb{R}^m \rightarrow [0,1]^T$	Neural network.
$\mathcal{L}(\cdot)$	An arbitrary loss function.
Smooth	Output smoothing function.
$\mathbb{Z}$	Adjacency matrix constructed from $\mathcal{M}$.
$\hat{\mathbb{Z}}$	SSA-elicited affinity matrix.
$\mathbb{A}$	Design matrix for geographical entities (descriptive geo feats).
Common	Function that determines whether two geographic entities in $\mathcal{M}$ are adjacent.
$\boldsymbol{Q}_{spec}$	Top k eigenvectors from $\boldsymbol{Q}$, selected based on largest eigenvalues in λ.
$\mathbf{q}_{label}$	The result of applying kMeans clustering to $\boldsymbol{Q}_{spec}$.
Enrich	Function that assigns values in $\boldsymbol{Q}_{spec}$ to an instance.
Θ	SSA procedure to produce $\hat{\mathbb{Z}}$.

2.2 Kaplan-Meier re-representation

To begin elaborating on the censored nature of our data, as we mentioned in the previous section, instance i has an event label $e^{(i)}$ and a discrete time of event occurrence $t^{(i)}$: provided this, the goal is to transform this two-valued representation to that of a *Kaplan-Meier survival curve* (KMSC) representation (Kaplan and Meier, 1958). A KMSC, simply put, each temporal unit $1,\ldots,T$ with a probability of the disease event $e^{(i)}$ *not* occurring at that particular temporal unit, dependent upon the probability of "not" event occurrence of the preceding temporal unit, for each instance i.

More formally, the KMSC re-representation is in the form of a vector, denoted $\mathbf{y}^{(i)} \in [0,1]^T$, where the index values $\tilde{t} \in \{1,\ldots,T\}$ express the temporal units and the indexed values $\mathbf{y}_{\tilde{t}}^{(i)}$ denote the respective probabilities.

Our KMSC re-representation scheme is originally outlined in Chi et al. (2007). To instantiate the vector $\mathbf{y}^{(i)}$, the following is conducted:

$$y_{\tilde{t}}^{(i)} = \begin{cases} 1 & \text{if } \tilde{t} < t^{(i)} \\ 0 & \text{if } \tilde{t} \geq t^{(i)} \,\&\, e^{(i)} = 1 \\ 1 - P(e_{\tilde{t}}^{(i)} = 1 \mid e_{\tilde{t}-1}^{(i)} = 0) & \text{if } \tilde{t} \geq t^{(i)} \,\&\, e^{(i)} = 0 \end{cases} \quad (1)$$

where $P(e_{\tilde{t}}^{(i)} = 1 \mid e_{\tilde{t}-1}^{(i)} = 0)$ denotes the conditional probability of event $e^{(i)}$ occurring at $\tilde{t}$ provided that $e^{(i)}$ has not occurred at $\tilde{t}-1$. As such, for patients whose CRC outcomes are known, $\mathbf{y}^{(i)}$ exhibits values that are strictly 0 and 1. On the other hand, a censored patient's vector exhibits estimation of survival probability beginning at the index position $\tilde{t} = t^{(i)}$; the ensuing values are conditional probability estimates.

2.3 Predicting individual KMSC

Our goal in this work is to induce an optimal hypothesis $\mathrm{g}^* \in \mathcal{G}$ of some [presently] arbitrarily defined hypothesis class $\mathcal{G}$, that is most apt at predicting instance-specific KMSCs. We formalise this problem as:

$$\mathrm{g}^* = arg\,min_{\mathrm{g}\in\mathcal{G}} \left\{ \mathcal{L}\left(\mathbf{y}^{(i)}, \mathrm{g}(\mathbf{x}^{(i)})\right) : i = 1,\ldots,n \right\} \quad (2)$$

where $\mathcal{L}(\cdot)$ expresses some loss function that measures the divergence between the predicted $\mathbf{y}^{(i)}$ (henceforth expressed $\hat{\mathbf{y}}^{(i)}$) and the known $\mathbf{y}^{(i)}$.

The hypothesis class $\mathcal{G}$ explored in this work is defined as both deep and shallow neural network architectures, the specifics of which are disclosed later in this section; we discuss the specific parameterisations employed across our experiments in the experiments section (Section 3). Deep neural network architectures are characterised by multiple hidden layers, and shallow architectures by a single hidden layer.

2.3.1 Output smoothing

Construction of a neural network model is accomplished in layer-wise fashion, where a particular layer is composed of nodes. The first layer in a neural network is designated as the input layer, which is proceeded by any number of so-called hidden layers, the last of which is connected to the output layer. The output layer is somewhat unique to our problem setting of predicting KMSCs. First, the predicted probability elicited from each of the $\tilde{t} = 1,\ldots,T$ output nodes are *ordered*. In other words, the output of $node_{\tilde{t}}^{out}$ is *ordered* before $node_{\tilde{t}+1}^{out}$ because $\tilde{t}$ is temporally occurs before $\tilde{t}+1$. Second, the ordered output probabilities of these nodes should be strictly decreasing: i.e., $output_{\tilde{t}}^{(i)} \geq output_{\tilde{t}+1}^{(i)}$. The reasoning behind this "strictly decreasing" expectation is intuitive: the probability of survival, of a disease or otherwise, even after disease recovery, is never expected to go up. The loss function $\mathcal{L}(\cdot)$ employed to induce mulitple-output networks, such as those in our problem setting, elicit a single loss value representing the loss across all nodes, meaning the desired strictly decreasing output

cannot be guaranteed. In light of this, we develop a smoothing operation, denoted $\text{Smooth}(\mathbf{output}^{(i)})$, formally expressed by

$$\hat{y}_{\tilde{t}+1}^{(i)} = \min\{output_{\tilde{t}}^{(i)}, output_{\tilde{t}+1}^{(i)}\} \text{ for } \tilde{t} = 1, \ldots, T \tag{3}$$

guaranteeing that the post-processed (i.e., smoothed) model output is strictly decreasing.

2.4 Geographic feature representation

Although our primary concern is to elicit a g* that produces the most accurate predictions, the novelty of the work is to:

1 Demonstrate that geographic-based feature representations enhance the quality of predictions.

2 Explore whether simple binary representations or a richer representations (defined shortly) produce more accurate predictions.

3 Quantify the extent to which these representations improve predictive quality.

The details of our experiments and data are elaborated on in the next section where we explore three different geographical representations: a simple binary representation (SBR), a rich representation based on spectral analysis (RR-SA), and a rich representation employing similarity-based spectral analysis (RR-SSA).

2.4.1 Simple binary representation

Our simple binary representation (SBR) is minimalist in nature, the procedure consisting only of (a) determination of the discrete geographic entity membership of instance i and (b) such membership being binarily re-represented (referred to as *one hot encoding*), thus eliciting a sparse vector-based encoding with a 1 in the indexical location referring to the geographic entity of which i is a member, and 0 s in the remaining positions.

To devise a formulation that is aptly generalisable we assume that the geographic features of instance i, expressed as $\mathbf{x}_{\mathbf{z}}^{(i)}$, are defined such that encoded values are capable of eliciting the discrete geographic unit of which i is a member (e.g., coordinates). For instance, we employ ZCTAs (zipcode tabulation area) as the discrete geographic unit in our experiments.

A formal procedure for eliciting discrete geographic unit membership can be expressed as

$$x_b^{(i)} = \Gamma(\mathbf{x}_{\mathbf{z}}^{(i)}, \mathcal{M}) \tag{4}$$

where the function $\Gamma(\cdot)$ performs a transformation on $\mathbf{x}_{\mathbf{z}}^{(i)}$, the geographic feature values of the instance, to some identification (ID) value, which we denote as $x_b^{(i)}$. This $x_b^{(i)}$ value denotes the unique, discrete geographic entity, belonging to map $\mathcal{M}$ (which we define momentarily), of which instance i is a member. The values represented by $\mathbf{x}_{\mathbf{z}}^{(i)}$, along with the information expressed in map $\mathcal{M}$, dictate the procedure used by $\Gamma(\cdot)$ to perform the transformation.

The specific $\mathbf{z}$ geographic features employed in this work are (latitude, longitude) coordinate pairs and, as such, we specify a definition (referred to as Definition 2.1) of map $\mathcal{M}$ using geography defined in terms of these coordinate pairs.

Definition 2.1: Define $\mathcal{M}$ to be a **map**, given by

$$\mathcal{M} = \{(key_l, value_l)\}_{l=1}^{p} \tag{5}$$

where key_l is the unique postal code of geographic unit l and $value_l$ is an ordered set of (lat,lon) coordinate pairs denoting the bounding geographic region of l.

We characterise map $\mathcal{M}$ as a continuous geographic region by

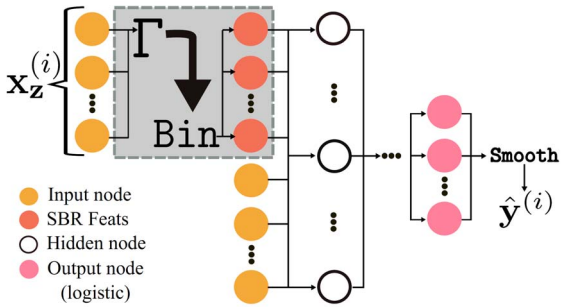
$$\left\{\forall \mathrm{l}\exists \mathrm{l}' : value_{\mathrm{l}}^{q} = value_{\mathrm{l}}^{j} \text{for} \mathrm{l}, \mathrm{l}' \in \{1,\ldots,p\} \& \mathrm{l} \neq \mathrm{l}'\right\} \tag{6}$$

where $value_{\mathrm{l}}^{q} = value_{\mathrm{l}}^{j} \triangleq (lat_{\mathrm{l}}^{q} = lat_{\mathrm{l}}^{j}) \cap (lon_{\mathrm{l}}^{q} = lon_{\mathrm{l}}^{j})$.

Provided our definition of map $\mathcal{M}$, expressed in Definition 2.1, we define $\Gamma(\cdot)$ to be a function that takes (lat,lon) coordinate pairs $\mathbf{x}_{\mathbf{z}}^{(i)}$ and determines whether the point is on the interior of each ZCTA. The zip code ID, corresponding to the ZCTA having the point $\mathbf{x}_{\mathbf{z}}^{(i)}$ on the interior, is subsequently initialised as the value of $x_b^{(i)}$ (i.e., $x_b^{(i)} = \mathrm{ZCTA}_{textID}$). Following this outlined $\Gamma(\cdot)$ procedure, and additional binarisation procedure, often referred to as one-hot encoding, which we denote Bin, is employed to produce a spare vector representation consisting of a single 1 in the position referencing the ZCTA of which instance i belongs, and 0 s in other positions.

Figure 2 illustrates the network architecture using the SBR methodology.

Figure 2 SBR neural network architecture



We expect that non-geographic representations will perform worse than representations employing the SBR representation

While models elicited from employing the SBR representation may enjoy some predictive performance improvement over hypotheses induced on strictly non-geogrpahic features, representations that consist of richer geographic encodings, capable of modelling the continuous nature of the geographic region of study promise to produce even better results.

2.4.2 Spectral analysis representation

To elicit richer geographical representations, we devise a spectral analysis approach, based on a well-known procedure referred to as spectral clustering. The method begins by first computing an adjacency matrix among the discrete geographic entities represented in $\mathcal{M}$. Subsequently, spectral analysis solves for the eigenvectors and eigenvalues of the adjacency representation, selecting the top k eigenvectors, based on the largest k eigenvalues. The elicited representation is $p \times k$ matrix, where the p rows refer to the p discrete geographic entities (i.e., a single row refers to one of the p entities). The k values that compose each row are used as geographic predictive input features.

To more formally relate this spectral analysis procedure, define $\mathbb{Z} = \text{Adj}(\mathcal{M})$ to be the affinity (i.e., adjacency, similarity) matrix, in which the l, v-th entry relates the geographic adjacency relationship among the l-th and v-th entities. We express this by

$$[\mathbb{Z}]_{l,v} = \begin{cases} 1 & \text{if Common}(values_l, values_v) = True \\ & \& l \neq v \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where the function Common$(\cdot)$ determines if $values_l$ and $values_v$ have a common element. Provided $\mathcal{M}$, related by Definition 2.1, Common$(\cdot)$ computes whether or not $values_l$ and $values_v$ share at least one coordinate pair.

Subsequently, spectral clustering is executed by performing k Means clustering, $\mathbf{q}_{label} = kMeans\left(\boldsymbol{Q}_{spec}\right)$, where the function $kMeans(\cdot)$ assigns one of the k cluster labels to each of the p entries of $\boldsymbol{Q}_{spec}$; where

$$\boldsymbol{Q}_{spec} = \text{Top}_k\left(\boldsymbol{Q}, \lambda\right). \tag{8}$$

The function $\text{Top}_k(\cdot)$ searches and finds the largest values in λ, selects the appropriate columns in the matrix $\boldsymbol{Q}$, and creates the matrix $\boldsymbol{Q}_{spec} \in \mathbb{R}^{k \times p}$. The matrix $\boldsymbol{Q}$, composed of eigenvectors, and corresponding vector λ, composed of eigenvalues, are obtained by solving the system of equations related by

$$\mathbb{Z}\boldsymbol{Q} = \lambda\boldsymbol{Q}. \tag{9}$$

Here, the columns of $\boldsymbol{Q}_{spec}$ are used as the k geographical features when inducing a hypothesis g – we refer to this use of the $\boldsymbol{Q}_{spec}$ matrix as spectral analysis. The labels, $\mathbf{q}_{label}$, obtained from application of the clustering procedure, referred to as spectral clustering, are used to visualise the elicited representations obtained from our experiments, related in the next section. Spectral analysis avoids making use of binarised label assignments of the spectral clustering procedure, instead using a sub-procedure, termed spectral analysis, which preserves cluster composition and is a richer (i.e., non-sparse) representation.

In Algorithm 1, we relate the spectral clustering process, differentiating spectral analysis from spectral clustering via red highlighting; omission of this line produces the

spectral analysis procedure. Simply put, spectral analysis is a sub-procedure of the spectral clustering process, yielded by omitting the clustering step.

Algorithm 1: Spectral clustering

1: Obtain adjacency matrix $\mathbb{Z}$ using (7).

2: Solve (9) for $\boldsymbol{Q}$ and λ.

3: Obtain $\boldsymbol{Q}_{spec}$ as outlined in (8).

4: Apply kMeans clustering to $\boldsymbol{Q}_{spec}$ to obtain q_{label}.

Finally, for a test instance $\mathbf{x}$, a process Enrich$(\mathbf{x}_{\mathbf{z}}, \mathcal{M}, \boldsymbol{Q}_{spec})$ is executed to obtain the appropriate k-valued column entry of $\boldsymbol{Q}_{spec}$ that is associated with the particular geographic entity that the test instance belongs to. Algorithm 2 outlines this procedure.

Algorithm 2: Enrich geographic features `Enrich`

Input: $\mathbf{x}_{\mathbf{z}}, \mathcal{M}, \boldsymbol{Q}_{spec}$

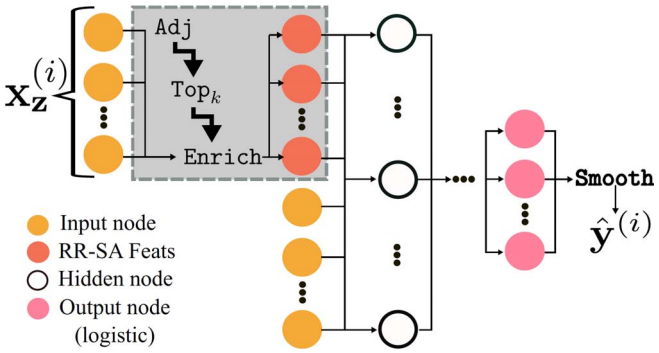
1: $x_b = \Gamma(\mathbf{x}_{\mathbf{z}}, \mathcal{M})$ From (4).

2: Using x_b find the l such that $x_b = key_l : l \in \{1, \ldots, p\}$.

Output: Return column vector $[\boldsymbol{Q}_{spec}]_l$.

The deep learning network architecture outlining the spectral analysis procedure in conjunction with learning a hypothesis, is depicted in Figure 3.[1]

Figure 3 RR-SA neural network architecture



2.4.3 Similarity-based spectral analysis representation

While the above spectral analysis-based approach produces richer geographic representations, the process is capable of leveraging only geographically relational information, as the input affinity matrix must be square (i.e., $p \times p$). It may, however, be beneficial to leverage encodings that are both relational and descriptive in nature, as geographically-descriptive features, such as population demographics and types of land-use, may further enrich the spectral analysis-elicited representations.

To allow for such input matrices we adjust our spectral analysis process to first calculate the pairwise similarity among geographic entities, thus producing a square affinity matrix on which the spectral analysis procedure can be performed.

More formally, recall the previously discussed adjacency matrix $\mathbb{Z}$ and define $\mathbb{A} \in \mathbb{R}^{p \times h}$ to be a geographic entity design matrix whose rows represent geographic entities and whose columns represent features. Additionally, $\mathbb{A}$ is constructed such that the l th row of $\mathbb{Z}$ and the l th row of $\mathbb{A}$ refer to the same entity.

Using $\mathbb{A}$ and $\mathbb{Z}$, along with a similarity measure, we devise two sub-representation methods from which a $p \times p$ affinity matrix can be derived.

The first sub-representation uses a single binary feature to indicate spatial adjacency between two geographic entities along with the geographically descriptive features of each entity to yield two vectors $\hat{\mathbf{z}}_l$ and $\hat{\mathbf{z}}_v$. These can be formally expressed as

$$\hat{\mathbf{z}}_l = [\mathbb{Z}_{l,v}, \mathbb{A}_l] \tag{10}$$

$$\hat{\mathbf{z}}_v = [\mathbb{Z}_{v,l}, \mathbb{A}_v]$$

where $[\cdot]$ represents the concatenation of a scalar or vector with another vector (in this case, it is scalar-vector concatenation). Also, note that $\mathbb{Z}_{l,v} = \mathbb{Z}_{v,l}$. We term this sub-representation method *SSA (bin)*.

The second sub-representation uses full geographic entity adjacency vectors instead of a definitive indicator of immediate spatial adjacency. This sub-representation method can be expressed by

$$\hat{\mathbf{z}}_l = [\mathbb{Z}_l, \mathbb{A}_l] \tag{11}$$

$$\hat{\mathbf{z}}_v = [\mathbb{Z}_v, \mathbb{A}_v]$$

where, in this case $[\cdot]$ indicates two vectors being concatenated. We term this sub-representation method *SSA (full)*.

Subsequently, using either SSA (bin) or SSA (full), the cosine similarity, denoted as $\phi(\cdot)$, between l and v is computed, thus producing an affinity matrix.

Algorithm 3, which denotes the procedure as $\boldsymbol{\Theta}$, fully discloses this process, using either SSA (bin) or SSA (full), while Figure 4 expresses the process in the context of the neural network architecture.

Algorithm 3: Sub-rep to affinity $\boldsymbol{\Theta}$

Input: $\mathbb{A}, \mathbb{Z}, \text{REP} \in \{\text{SSA(bin)}, \text{SSA(full)}\}$

1: $\hat{\mathbb{Z}} \leftarrow \mathbf{0}^{p \times p}$

2: **for** $l = 1, \ldots, p-1$ **do**

3: **for** $v = l+1, \ldots, p-1$ **do**

4: **if** REP = SSA (bin) **then**

5: Define $\hat{\mathbf{z}}_l$ and $\hat{\mathbf{z}}_v$ according to (10).

6: **else**

7: Define $\hat{\mathbf{z}}_l$ and $\hat{\mathbf{z}}_v$ according to (11).

8: **end if**

9: $\hat{\mathbb{Z}}[i,j] \leftarrow \phi(\hat{\mathbf{z}}_l, \hat{\mathbf{z}}_v)$

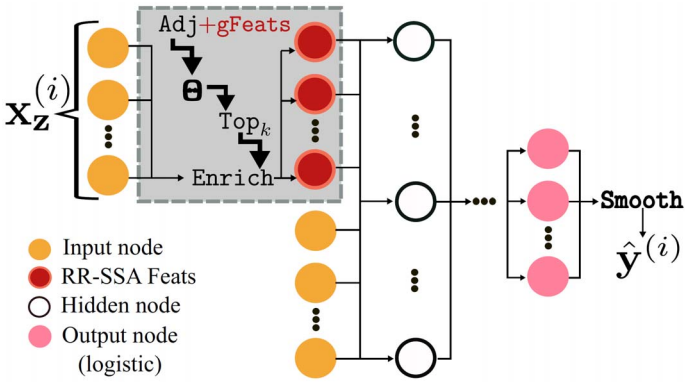
10: $\hat{\mathbb{Z}}[j,i] \leftarrow \phi(\hat{\mathbf{z}}_l, \hat{\mathbf{z}}_v)$

11: **end for**

12: **end for**

Output: Return $\hat{\mathbb{Z}}$

Figure 4 RR-SSA neural network architecture



3 Predicting colorectal cancer survival

We begin this section with an in-depth disclosure of the data employed in our experiments, subsequently outlining the technicalities involved in undertaking these experiments. Finally, we provide a discussion of the results elicited from performing these experiments by comparing the average predicted survival curve of each method against the average actual survival curve, leveraging a devised measure, referred to as *area between the curves* (ABC), discussed further on in this section.

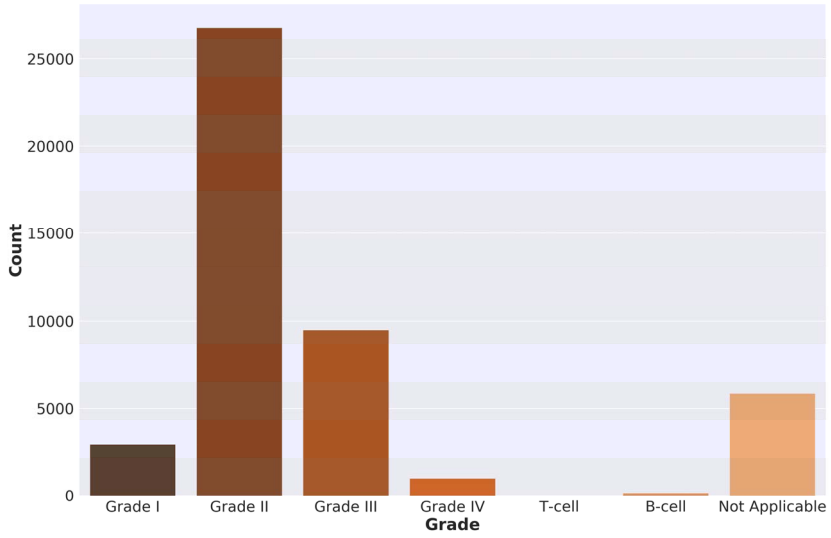
3.1 Colorectal cancer survival data for the state of Iowa

Our data were provided by the Iowa Cancer Registry (ICR), State Health Registry of Iowa (SHRI), and the Iowa Department of Public Health (IDPH). Each instance represents a patient who has been diagnosed with colorectal cancer and whose residence at the time of diagnosis is in the state of Iowa. The dataset consists of $n = 46{,}116$ patients and, initially, $m = 71$ features. After removing identifiers and features having a large number of instances with missing values (% missing > 50%), we were left with $m = 26$ distinct features (including unprocessed geographic coordinates). After binarising discrete features, $m = 386$ (excluding geographic features). When using SBR geographical re-representation, $m = 1364$ (386 non-geographic features and $p = 978$

binarised geographic features), and $m = 386 + k$ when using the RR-SA geographic representation (where k is parameterised and therefore user-dependent). When the Kaplan-Meier re-representation is applied to the dataset, we obtain $\mathbf{y}^{(i)}$ vectors having $T = 53$ elements, where each element represents the patient's current vital status (alive = 1 or dead = 0), or a probability of survival when an instance becomes censored, as described by (1). Each $\tilde{t} \in \{1, \ldots, 53\}$ represents six months.

The 24 distinct non-geographic features pertain to various patient-specific characteristics, which can be categorised as *disease-based* and *demographic-based*. Disease-based features include tumour grade, tumour histology and tumour marker; we show a histogram of tumour grade in Figure 5. Demographic-based features include marital status, race, and age at diagnosis; we show a histogram of age at diagnosis in Figure 6. These selected features (age and tumour grade) have been shown to be indicative of not receiving timely cancer treatment (Ward et al., 2013), which we believe will help in predicting cancer survival, although analysis of such factors is beyond the scope of this work.

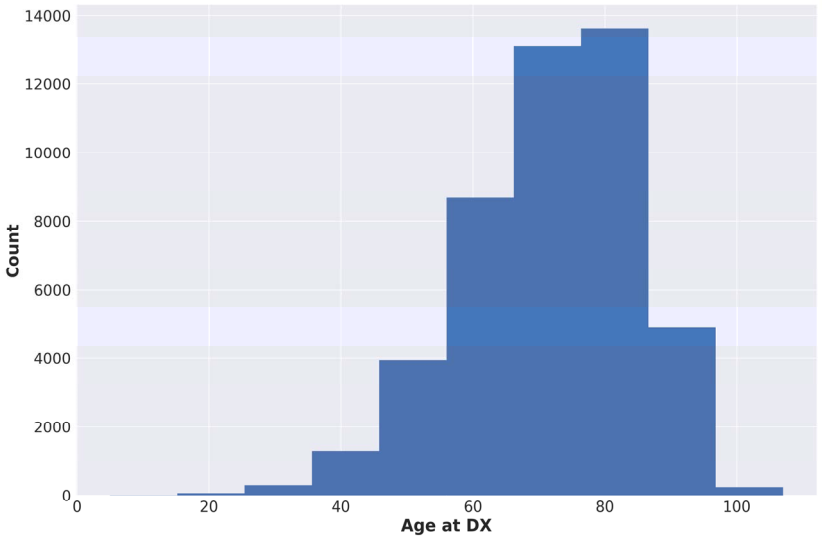
Figure 5 Tumour grade at diagnosis for patients in the state of Iowa: years 1989 to 2013



3.2 *Geographically-descriptive features for the state of Iowa*

We obtain geographically-descriptive features for the state of Iowa at the ZCTA-level of spatial granularity from the US Census Bureau's American FactFinder 2 website. Three different geographically-descriptive features were obtained for each of the 978 ZCTAS in Iowa: population age demographics, land type, and median household income. Population age demographics and land type are categorical features and were represented in terms of proportional bins (e.g., "percentage of population aged 0–5 years").

Figure 6 Age of colorectal cancer diagnosis for patients in the state of Iowa: years 1989 to 2013



3.3 Predictive setting, paramaterisation and results

As outlined in the introduction, we wish to address the following:

1 On average, can colorectal cancer survival curves be reasonably predicted for patients in the state of Iowa?

2 Do geographic features improve the quality of predicted colorectal cancer survival curves for patients in the state of Iowa?

3 Do richer geographical feature representations improve predictive performance more than simpler representations?

4 Can predictive performance be further improved by altering the RR-SA procedure to accommodate adjacency-descriptive geographical feature pairings (i.e., RR-SSA)?

5 Which RR-SSA representation improves predictive performance the most: binary (bin) or full?

To such an end, we propose to use tenfold validation where, for each fold, we find a g* for each of the following types of model:

1 A model constructed using no geographical features (No Geo).

2 A model constructed using SBR-derived geographical features, as outlined by Figure 2 (SBR).

3 Models constructed using RR-SA-derived geographical features, as outlined by Figure 3, where the values $k = 10, 20, 30, 40$ will be explored (RR-SA).

4 Models construct using RR-SSA-derived geographical features, as outlined by Figure 4, using the *binary-based* adjacency representation, where the values $k = 10, 20, 30, 40$ will be explored (RR-SSA (bin)).

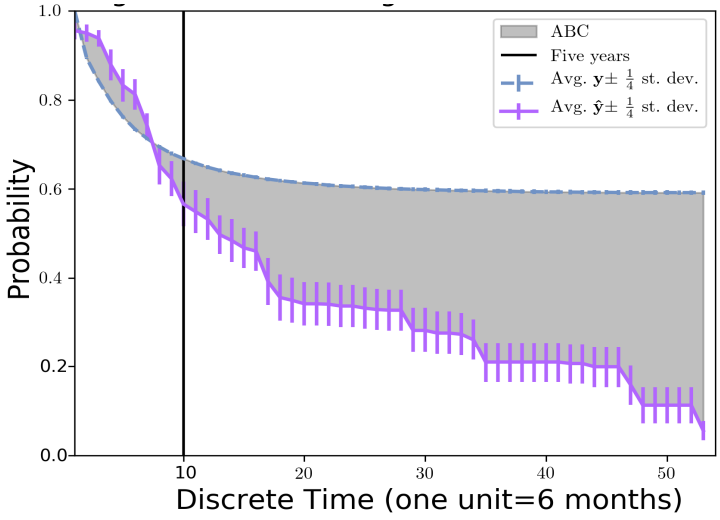
5 Models construct using RR-SSA-derived geographical features, as outlined by Figure 4, using the *full* adjacency representation, where the values $k = 10, 20, 30, 40$ will be explored (RR-SSA (bin)).

We then assess predictive performance by computing each model's average survival curve prediction on the test set, taken over tenfolds, as compared to the actual average survival curve, taken over all $\mathbf{y}^{(i)}$, using a measure termed *area between curves* (ABC) that measures the area-wise disparity between the actual and predicted curves (Lash et al., 2017b).

3.3.1 Model parameterisation

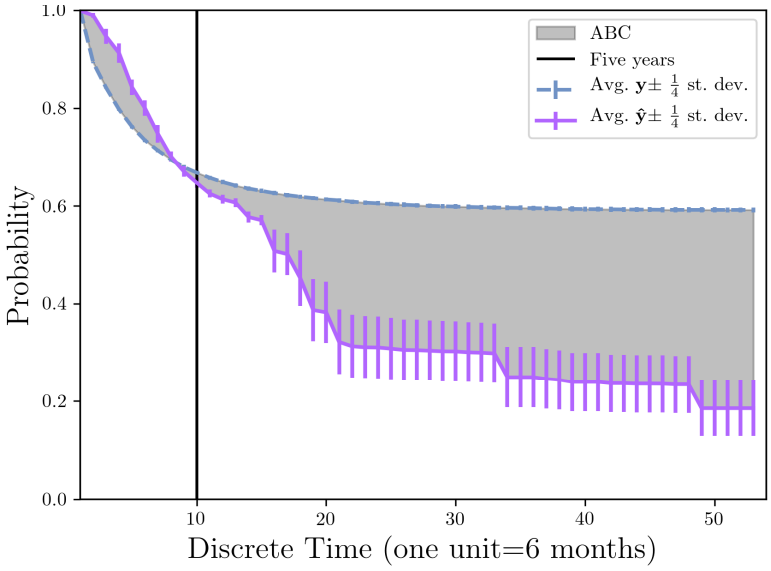
Our models are constructed using Tensorflow, employing fully connected layers, trained using sigmoidal cross entropy as the loss function $\mathcal{L}(\cdot)$. The logistic activation function is used for all nodes. Each model is trained using a maximum of 2500 epochs with batch size ranging from 5% to 20%. While the connectedness of the architecture, activation function, epochs, and batch size are all tunable parameters, we elect to focus on finding the optimal number of hidden layers and corresponding hidden nodes for each layer (note that epochs of 1000, 1500, 2000, and 2500 were explored). Table A1, in Appendix A, shows the average optimal architecture for each of the models, taken over the tenfolds.

Figure 7 Actual vs. predicted: *k* (when specified) denotes the parameterised *k* for spectral analysis, and ABC represents the *area between curves*

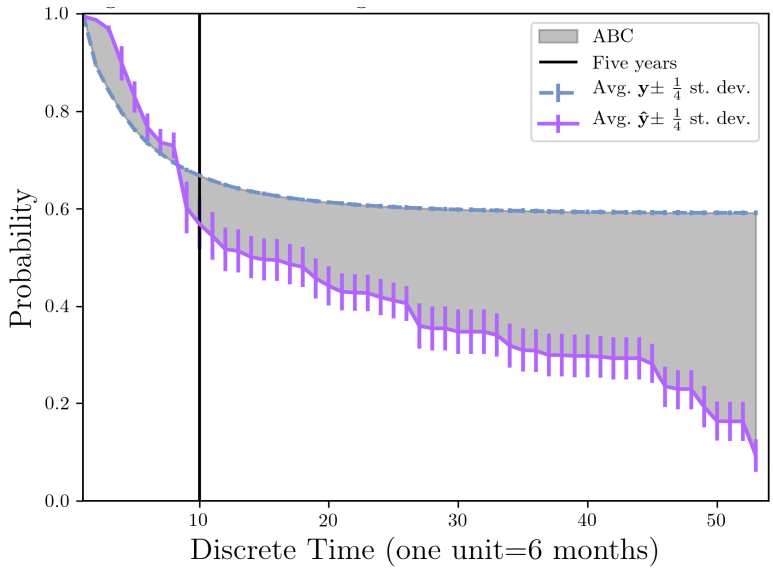


(a) No geo feats (ABC = 14.32)

Figure 7 Actual vs. predicted: k (when specified) denotes the parameterised k for spectral analysis, and ABC represents the *area between curves* (continued)

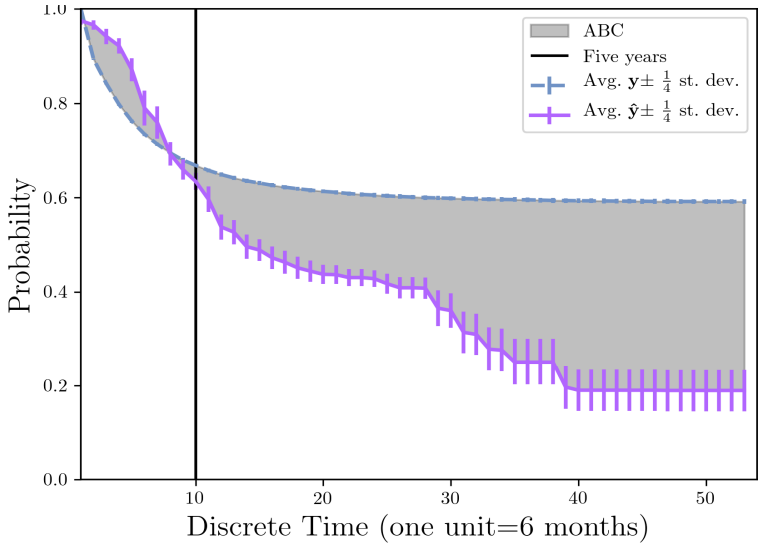


(b) SBR (ABC = 12.60)

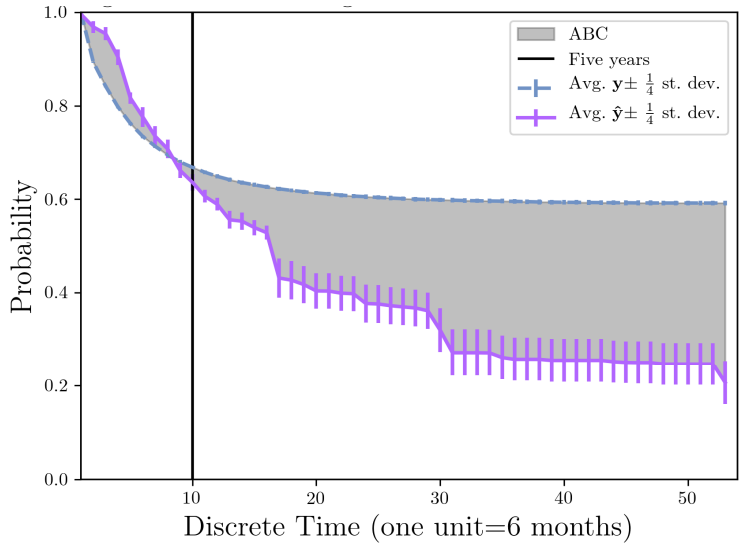


(c) RR-SA, $k = 10$ (ABC = 11.41)

Figure 7 Actual vs. predicted: k (when specified) denotes the parameterised k for spectral analysis, and ABC represents the *area between curves* (continued)

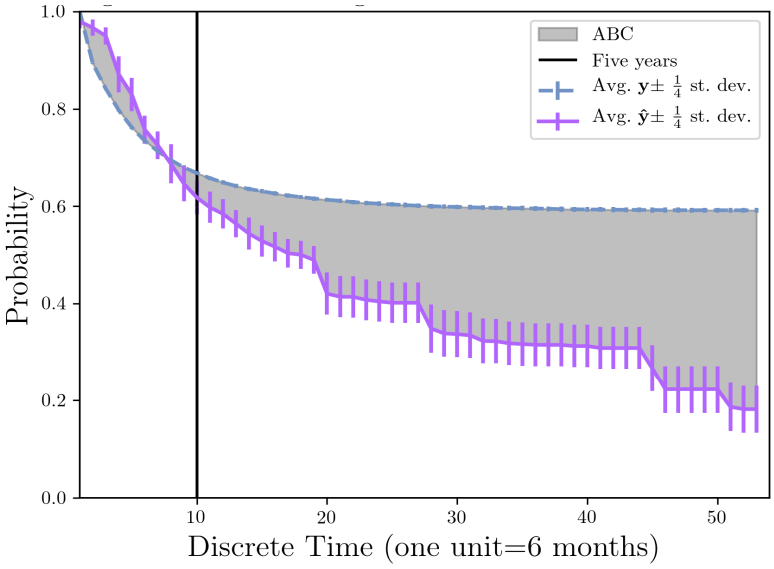


(d) RR-SA, $k = 20$ (ABC = 12.31)

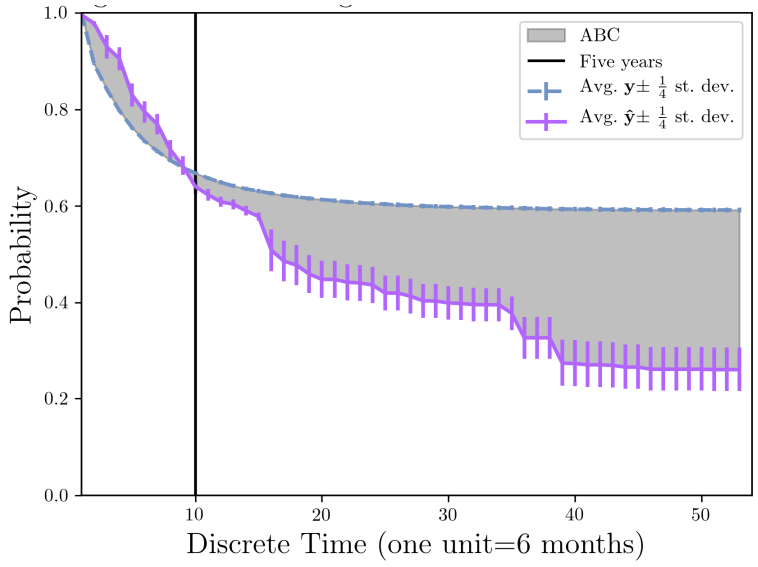


(e) RR-SA, $k = 30$ (ABC = 11.65)

Figure 7 Actual vs. predicted: k (when specified) denotes the parameterised k for spectral analysis, and ABC represents the *area between curves* (continued)

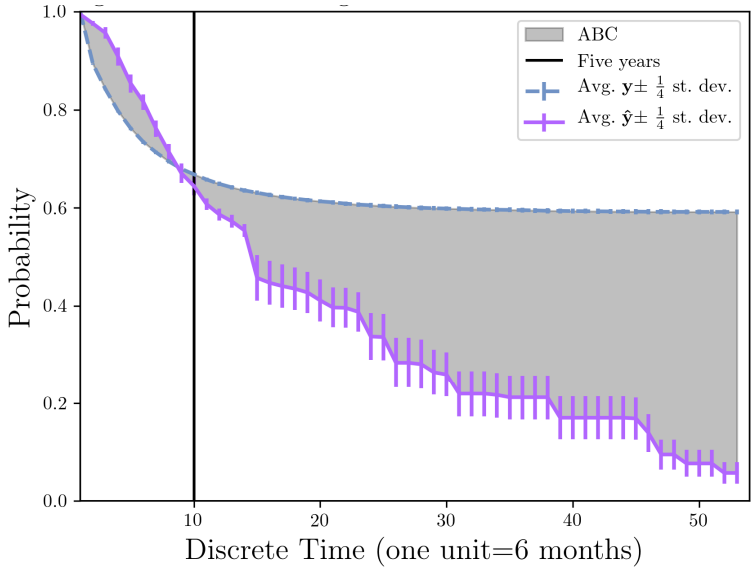


(f) RR-SA, $k = 40$ (ABC = 10.77)

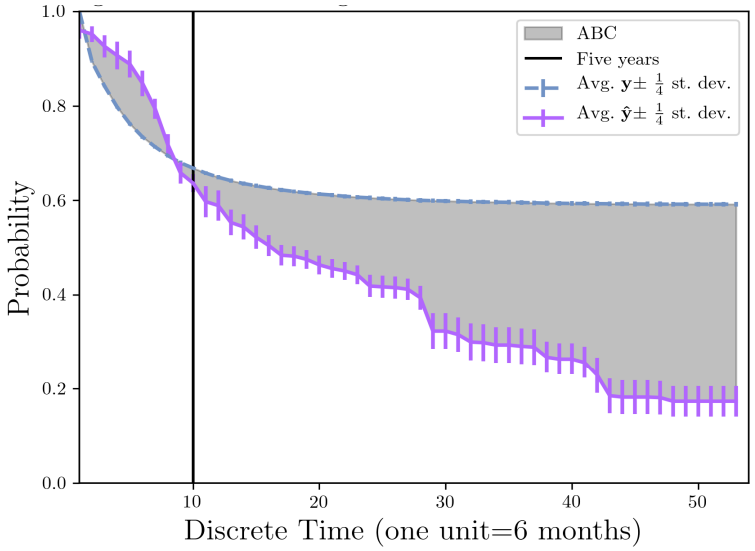


(g) RR-SSA (bin), $k = 10$ (ABC = 9.805)

Figure 7 Actual vs. predicted: k (when specified) denotes the parameterised k for spectral analysis, and ABC represents the *area between curves* (continued)

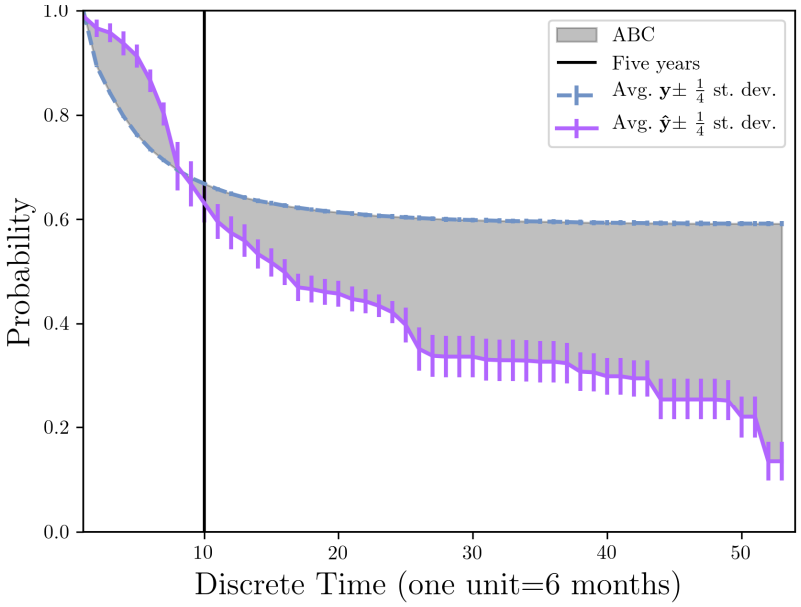


(h) RR-SSA (bin), $k = 20$ (ABC = 14.555)

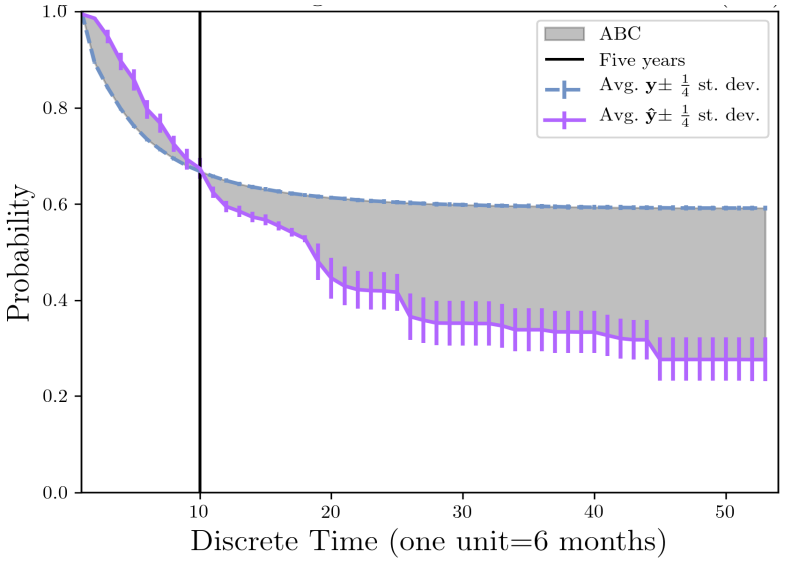


(i) RR-SSA (bin), $k = 30$ (ABC = 11.850)

Figure 7 Actual vs. predicted: k (when specified) denotes the parameterised k for spectral analysis, and ABC represents the *area between curves* (continued)

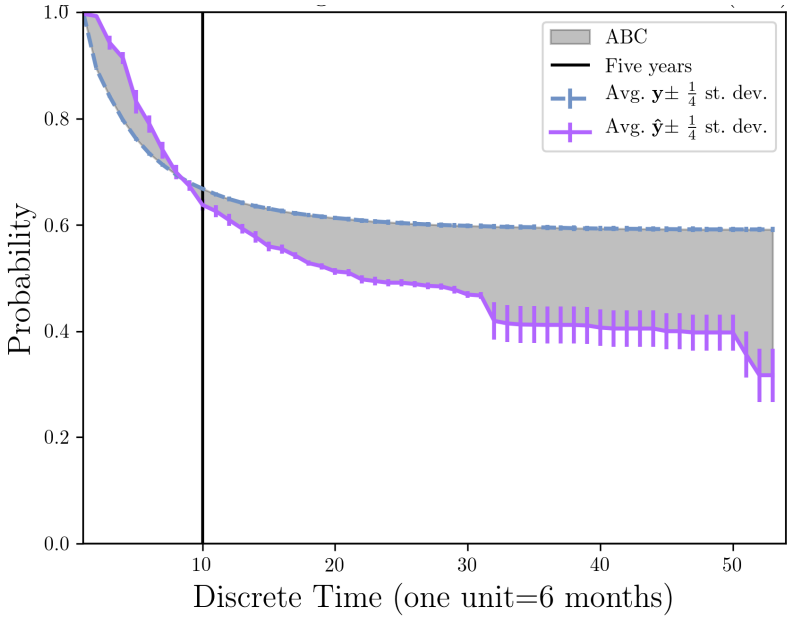


(j) RR-SSA (bin), $k = 40$ (ABC = 11.205)

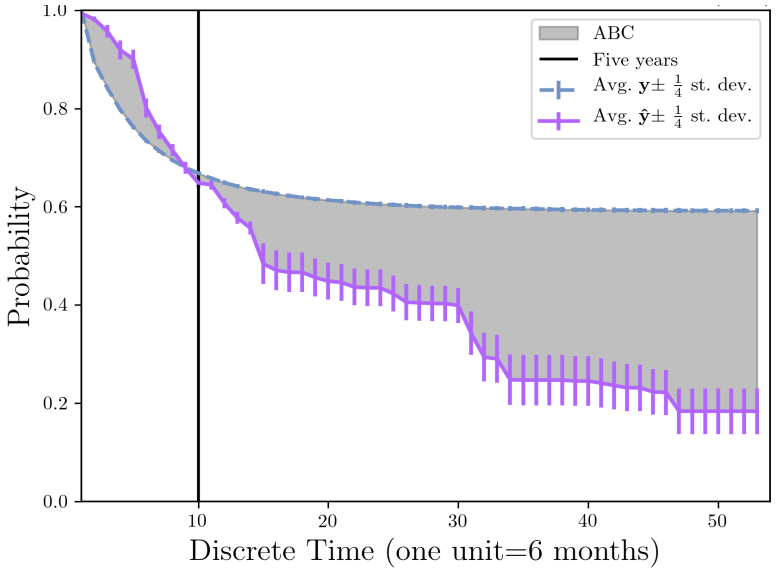


(k) RR-SSA (full), $k = 10$ (ABC = 9.820)

Figure 7 Actual vs. predicted: k (when specified) denotes the parameterised k for spectral analysis, and ABC represents the *area between curves* (continued)

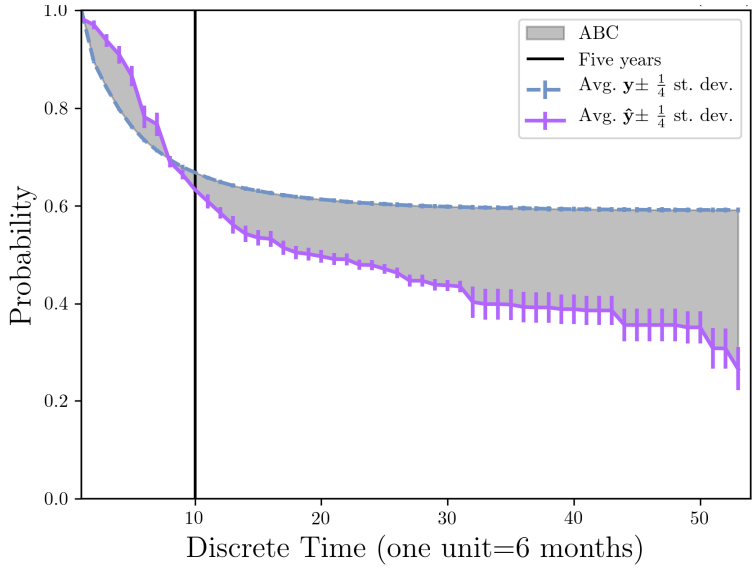


(l) RR-SSA (full), $k = 20$ (ABC = 6.657)



(m) RR-SSA (full), $k = 30$ (ABC = 11.724)

Figure 7 Actual vs. predicted: k (when specified) denotes the parameterised k for spectral analysis, and ABC represents the *area between curves* (continued)



(n) RR-SSA (full), $k = 40$ (ABC = 7.818)

3.3.2 *Average actual vs. average predicted survival*

The results comparing the average actual survival curve against the average predicted survival curve, by model, are presented in Figure 7. Henceforth, these curves will simply be referred to as *actual* and *predicted*. In these figures we also shade the region between the actual and predicted curves and provide a value representing the total area covered by this region. We will use this measure, developed in Lash et al. (2017b), referred to as *area between the curves* (ABC for short), as a means of comparing the predictive quality of the 14 different models (where lower ABC is better).

Comparing Figure 7a with Figures 7b through 7n we first see that the addition of geographical features has uniformly improved the quality of the predictions, on average, as can be observed visually and by comparing ABC values. That is, with the exception of RR-SSA (bin) $k = 20$, which suggests that it is important to tune the spectral analysis k value when using such representations.

Secondly, comparing Figure 7b with Figures 7c through 7f, we observe that models using richer geographical representations (RR-SA) perform better (7c–7f) than a model trained using a simple representation (7b). Furthermore, employing SSA-based representations yield even better improvements over SBR, depending on the parameterised value of k.

However, there are also RR-SA model performance differences depending on the parameterised k value. Interestingly, there seems to exist a non-linear relationship between k and performance, with $k = 10$ outperforming $k = 20$, and $k = 30$ outperforming $k = 10$; $k = 40$ performs the best out of all models. We believe this nonlinear relationship may be accounted for by the fact that higher values of k lead to

more localised models, yet can also produce sparse, disjointed clusters. This point is supported by our clustering visualisations reported in Figure 8 and discussed in Section 3.3.3. These nonlinear response observations can also be extended to RR-SSA (bin) and RR-SSA (full).

Comparing RR-SA with RR-SSA representations, we can see even greater improvement in our predictions, on average. In fact, by employing RR-SSA (full), we achieve a 38.2% relative improvement in ABC value when comparing the best RR-SSA (full) result ($k = 20$) with the best RR-SA result ($k = 40$). Interestingly, and perhaps not entirely unexpectedly, RR-SSA (bin) obtained less predictive improvement when compared with RR-SSA (full), but is able to improve upon the RR-SA result.

Curiously, however, depending upon the parameterised k value, RR-SSA (bin) performs worse than models induced without geographical features and those induced using SBR. We conjecture that this may be attributable to the overly simple representation of geographic adjacency used in the sub-representation method of RR-SSA (bin). This is a reasonable conclusion as we can see that using a "fuller" representation (i.e., RR-SSA (full)) of adjacency produces uniformly improved results.

In examining the different predicted survival curves we have a few observations, summarised as follows. First, we observe that predictive performance increases are mostly realised after the five-year mark. This is, on one hand, intuitive because predicting survival at times closer to the diagnosis is easier than predicting survival at later times. On the other hand, noticeable deviation of the predicted curves uniformly occurs across all models at or around this five-year mark. Therefore, model improvement wrought by using richer geographical representations is realised, by-in-large, at times beyond the five-year mark. Explanation as to *why* such a deviation is present in all models requires further investigation beyond the scope of this work.

In summary, we find that

1 On average, colorectal cancer survival curves can be reasonably predicted for patients in the state of Iowa.

2 Geographic features do improve the quality of predicted colorectal cancer survival curves for patients in the state of Iowa by 53.5% (on average) (comparing models induced without geographic features with RR-SSA (full) $k = 20$).

3 On average, RR-SA feature representations improve predictive performance by 15% over simple representations (SBR) and RR-SSA improve predictive performance by 47.2% over SBR.

4 On average, RR-SSA feature representations improve predictive performance by 38.2% over RR-SA representations (comparing RR-SSA (full) $k = 20$ to RR-SA $k = 40$).

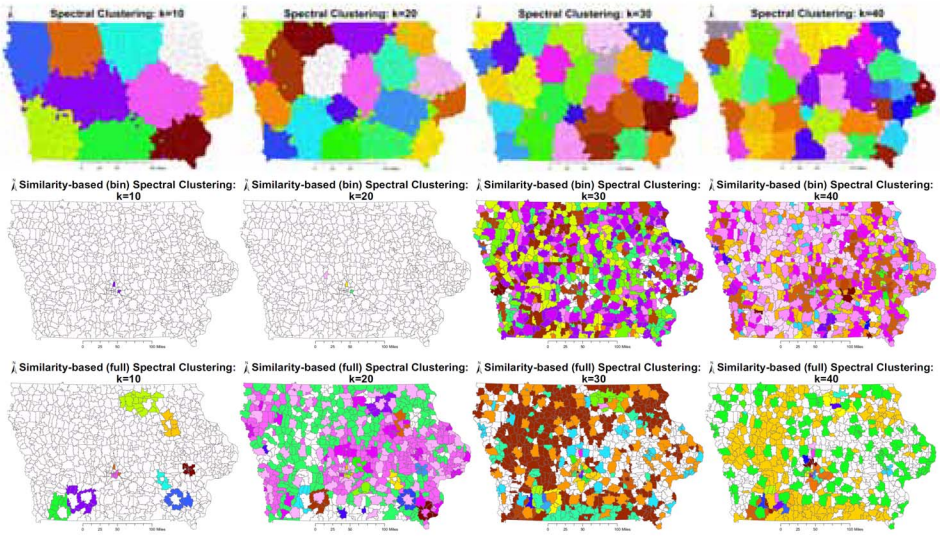
5 On average, RR-SSA (full) feature representations improve predictive performance by 32.1% (comparing RR-SSA (full) $k = 20$ to RR-SSA (bin) $k = 10$).

3.3.3 Visualising geographic cluster assignment

Next, we briefly discuss the results of visualising cluster assignment for $k = 10, 20, 30, 40$ for RR-SA, RR-SSA (bin), and RR-SSA (full). These results can be observed in Figure 8, where each unique colour represents a single cluster.

For RR-SA (row 1), we first note that as k increases, the elicited geographic regions become more precise, yet maintain geographic continuity. However, we secondly observe that some ZCTAs are not adjacent to any other ZCTA having the same cluster assignment. This disjointedness stems from the use of an adjacency representation of the affinity matrix on which spectral clustering is performed and is not unexpected. As k increases it appears that the number of disjointed ZCTAs also increases. However, we see that the number of continuous regions also increases. In other words, while disjointedness seems to increase with k, the desired result of more localised continuous geographical regions is still achieved. Interestingly, when $k = 40$, larger Iowa cities such as Des Moines (central Iowa) and Iowa City (central-eastern Iowa) begin to emerge.

Figure 8 Spectral clustering results for $k = 10, 20, 30, 40$, where colour denotes cluster membership. Row one represents RR-SA results, row two represents RR-SSA (bin) results, and row three represents RR-SSA (full) clustering results



In examining row 2 of Figure 8, representing the RR-SSA (bin) results we see entirely different geographic clusterings than that of RR-SA. First, we find that for smaller values of k ($k = 10,20$), cluster membership is very skewed, with a single cluster dominating the majority of the state, and the remaining cluster assignments being composed of single ZCTAs. These single-ZCTA clusters are found in Des Moines area, the largest urban area of Iowa. As k is increased (i.e., $k = 30$, 40), rural areas begin to decompose into cluster subsets – i.e., as the representation is allowed to become more specific (by increasing k), rural areas begin to become distinguished between. Urban areas, such as Des Moines and Iowa City, are also ascribed membership to clusters composed of fewer geographic entities.

Looking at row 3 of Figure 8, which constitutes the cluster results obtained from RR-SSA (full) models, we observe different clustering results from that of the previous two models. First, we can see that clusters often form "ring-like" patterns (this is particularly observable for $k = 10$), which is a particularly interesting artefact of this representation. Secondly, juxtaposing these results, with that of the previous two rows (i.e., RR-SA and RR-SSA (bin)), we observe that this representation is somewhat of a

"compromise" between RR-SA and RR-SSA (full) in the sense that RR-SA produces mostly geographically contiguous clusterings and RR-SSA (bin) produces more geographically disparate clusterings. This is not unexpected, as the sub-representation method of RR-SSA (full) employs the full adjacency representation used in RR-SA, which is not found in RR-SSA (bin). Interestingly, RR-SSA (full) has also discovered urban areas such as Des Moines and Iowa City, but does so at smaller values of k than RR-SSA (bin) (e.g., RR-SSA (bin) $k = 10$ is only able to discern areas around Des Moines, whereas RR-SSA (full) is able to discern Iowa City, Des Moines, Waterloo/Cedar Falls, Mason City, etc.). Finally, as k is increased we observe that the representation is becomes more specific in terms of both urban and rural areas up to $k = 30$. When $k = 40$ we observe that the clusterings are more disparate, where there appear to be approximately three different rural areas distinguished between (yellow, green, and white), and where urban ZCTAs are assigned to their own unique cluster. This may suggest that urban areas are much more heterogeneous than are rural areas.

4 Related work

The topics related to and discussed throughout this work can best be categorised as *disease and survival curve prediction* and *geographic-based predictions and representation*.

There are many past works involving the prediction of diseases. These can be viewed as classification-based (Belciug and Gorunescu, 2013; Belciug, 2010; Gupta et al., 2011; Khosravi et al., 2015; Ojha and Goel, 2017; Puddu and Menotti, 2012; Sandhu et al., 2015) and survival-based (Chi et al., 2007; Cox, 1992; Gupta et al., 2011; Katzman et al., 2016; Samundeeswari and Saranya, 2016; Sharma et al., 2017). The focus of this work was on survival curve predictions. Such works can be examined by method, which include Cox proportional hazards model (CPH) (Cox, 1992), which has been historically used to make such predictions, decision trees (Sharma et al., 2017), and neural network-based models (Chi et al., 2007; Gupta et al., 2011; Katzman et al., 2016; Samundeeswari and Saranya, 2016), which are a more recent development. However, as De Laurentiis and Ravdin (1994) point out, CPH has several caveats as compared to neural network-based approaches, including the naivety of the proportional hazards assumption and inability to capture nonlinear feature interactions. Furthermore, decision trees are constructed using greedy methodology and do not have the architectural benefits of neural networks. Hence, this work employed neural networks.

There are also many works focusing on *geographic-based prediction and representation*. These works focus on incorporating geographical features into the predictive process. One method of representing geography is by fine grain lattice (i.e., grid) (Khezerlou et al., 2017; Lash et al., 2017a; Yuan et al., 2017). Such methods are akin to our SBR representation and suffer from the same shortcomings. Spatially adaptive filters (Tiwari and Rushton, 2005), which can tie a single feature to geography when creating $\mathcal{M}$, which may be beneficial when the selected feature is particularly indicative of survival. This method would, however, still produce a binary feature representation, having the accompanying shortcomings discussed when disclosing SBR. Spectral clustering has been used to cluster both social networks (White and Smyth, 2005) and for representing geo-spatial features (Frias-Martinez and Frias-Martinez, 2014; Van Gennip et al., 2013), as in this work, and produces a rich (i.e., non-sparse) vector of features.

5 Conclusions and future work

In this work we explored the use of four different geographical feature representations – a simple binary representation (SBR) and a rich representation based on spectral analysis (which we term spectral analysis and methodologically refer to as RR-SA), and two representations based on similarity-based spectral analysis (RR-SSA) – to predict colorectal cancer survival curves for patients in the state of Iowa. We show that (a) survival curves can be reasonably estimated, although predictive performance deviates near the five-year survival mark, (b) the use of geographical features generally lead to better predictions, (c) RR-SA trained models outperform those trained using SBR, (d) RR-SSA induced models, generally, outperform RR-SA models, and (e) RR-SSA (full) representations outperform RR-SSA (bin) representations. Future work will involve exploration of different geographical representations, particularly those learned in conjunction with g*. Additionally, continued exploration of domains and scenarios in which SBR, RR-SA, and RR-SSA geographic representations provide benefit should be undertaken.

Acknowledgements

The authors would like to thank the Iowa Cancer Registry, State Health Registry of Iowa, and the Iowa Department of Public Health for the data. The authors would also like to thank Gary Hulett and Jason Brubaker for their help in dataset construction and Prakash Nadkarni for his help with both data acquisition and the IRB process.

References

Belciug, S. (2010) 'A two stage decision model for breast cancer detection', *Annals of the University of Craiova-Mathematics and Computer Science Series*, Vol. 37, No. 2, pp.27–37.

Belciug, S. and Gorunescu, F. (2013) 'A hybrid neural network/genetic algorithm applied to breast cancer detection and recurrence', *Expert Systems*, Vol. 30, No. 3, pp.243–254.

Chi, C-L., Street, W.N. and Wolberg, W.H. (2007) 'Application of artificial neural network-based survival analysis on two breast cancer datasets', *AMIA Annual Symposium Proceedings*, Vol. 2007, p.130.

Cox, D.R. (1992) 'Regression models and life-tables', *Breakthroughs in Statistics*, Springer, pp.527–541.

De Laurentiis, M. and Ravdin, P.M. (1994) 'A technique for using neural network analysis to perform survival analysis of censored data', *Cancer Letters*, Vol. 77, Nos. 2/3, pp.127–138.

Frias-Martinez, V. and Frias-Martinez, E. (2014) 'Spectral clustering for sensing urban land use using twitter activity', *Engineering Applications of Artificial Intelligence*, Vol. 35, pp.237–245.

Gupta, S., Kumar, D. and Sharma, A. (2011) 'Data mining classification techniques applied for breast cancer diagnosis and prognosis', *Indian Journal of Computer Science and Engineering*, Vol. 2, No. 2, pp.188–195.

Kaplan, E.L. and Meier, P. (1958) 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association*, Vol. 53, No. 282, pp.457–481.

Katzman, J., Shaham, U., Bates, J., Cloninger, A., Jiang, T. and Kluger, Y. (2016) 'Deep survival: a deep cox proportional hazards network', *arXiv preprint arXiv:1606.00931*.

Khezerlou, A.V., Zhou, X., Li, L., Shafiq, Z., Liu, A.X. and Zhang, F. (2017) 'A traffic flow approach to early detection of gathering events: comprehensive results', *ACM Transactions on Intelligent Systems and Technology*, Vol. 8, No. 6, pp.74:1–74:24.

Khosravi, B., Pourahmad, S., Bahreini, A., Nikeghbalian, S. and Mehrdad, G. (2015) 'Five years survival of patients after liver transplantation and its effective factors by neural network and cox proportional hazard regression models', *Hepatitis Monthly*, Vol. 15, No. 9.

Lash, M.T., Slater, J., Polgreen, P.M. and Segre, A.M. (2017a) 'A large-scale exploration of factors affecting hand hygiene compliance using linear predictive models', *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*, pp.66–73.

Lash, M.T., Sun, Y., Zhou, X., Lynch, C.F. and Street, W.N. (2017b) 'Learning rich geographical representations: predicting colorectal cancer survival in the state of Iowa', *Bioinformatics and Biomedicine (BIBM'17), 2017 IEEE International Conference on*, IEEE, pp.778–785.

Ojha, U. and Goel, S. (2017) 'A study on prediction of breast cancer recurrence using data mining techniques', *Cloud Computing, Data Science & Engineering-Confluence, 2017 7th International Conference on*, IEEE, pp.527–530.

Puddu, P.E. and Menotti, A. (2012) 'Artificial neural networks versus proportional hazards cox models to predict 45-year all-cause mortality in the Italian rural areas of the seven countries study', *BMC Medical Research Methodology*, Vol. 12, No. 1, p.100.

Samundeeswari, E.S. and Saranya, P.K. (2016) 'An artificial neural network model for prediction of survival time of breast cancer dataset', *International Journal of Research in Engineering and Applied Sciences*, Vol. 6, No. 1, pp.161–168.

Sandhu, I.K., Nair, M., Shukla, H. and Sandhu, S.S. (2015) 'Artificial neural network: As emerging diagnostic tool for breast cancer', *International Journal of Pharmacy and Biological Sciences*, Vol. 5, No. 3, pp.29–41.

Sharma, A., Karthik, G.S., Mittal, N., Sindhu, V.L. and Pradeep, K.R. (2017) 'A survey on predictive analysis of cancer survivability rate using machine learning algorithm', *7th International Conference on Recent Trends in Engineering, Science, and Management*, pp.271–278.

Tiwari, C. and Rushton, G. (2005) 'Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa', *Developments in Spatial Data Handling, Proceedings of the 11th International Symposium on Spatial Data Handling*, Springer, Berlin, Heidelberg, pp.665–676.

Van Gennip, Y., Hunter, B., Ahn, R., Elliott, P., Luh, K., Halvorson, M. et al. (2013) 'Community detection using spectral clustering on sparse geosocial data', *SIAM Journal on Applied Mathematics*, Vol. 73, No. 1, pp.67–83.

Wan, N., Zhan, F.B., Zou, B. and Wilson, J.G. (2013) 'Spatial access to health care services and disparities in colorectal cancer stage at diagnosis in Texas', *The Professional Geographer*, Vol. 65, No. 3, pp.527–541.

Ward, M.M., Ullrich, F., Matthews, K., Rushton, G., Goldstein, M.A., Bajorin, D.F., Hanley, A. and Lynch, C.F. (2013) 'Who does not receive treatment for cancer?' *Journal of Oncology Practice*, Vol. 9, No. 1, pp.20–26.

White, S. and Smyth, P. (2005) 'A spectral clustering approach to finding communities in graphs', *Proceedings of the 2005 SIAM international conference on data mining*, pp.274–285.

Yuan, Z., Zhou, X., Yang, T., Tamerius, J. and Mantilla, R. (2017) 'Predicting traffic accidents through heterogeneous urban data: a case study', *6th International Workshop on Urban Computing (UrbComp 2017)*.

Zhang, R., Li, N., Yang, X. and Huang, Y. (2015) 'Data mining technology and its application in diagnosis and treatment of clinical malignant tumor', *Journal of Medical Informatics*, pp.50–54.

Note

1 As previously mentioned, $\mathbf{x}_{\mathbf{z}}^{(i)}$ denote (latitude, longitude) coordinate pairs.

Appendix A

In Table A1 we can see that, on average, the optimal architecture is relatively comparable among all models with the exception of SBR (and to a degree RR-SA, $k = 20$). First, this suggests that the use of RR-SS and RR-SSA features do not affect the architectural complexity of the model. However, SBR seems to significantly increase such complexity. This is somewhat expected, as SBR is represented as a large, sparse vector, which can be contrasted with the comparatively small vector of RR-SA and RR-SSA.

Table A1 Average optimal architecture by model over the tenfolds (e.g., No geo had 1.5 hidden layers, on average, where the first layer had 83 nodes, on average, and the second layer had 30 nodes, on average)

Model	*Average optimal architecture*
No Geo	1.5:[83,30]
SBR	1.9:[260,122]
RR-SA, $k = 10$	1.5:[82,36]
RR-SA, $k = 20$	1.5:[102,44]
RR-SA, $k = 30$	1.6:[87,45]
RR-SA, $k = 40$	1.5:[80,44]
RR-SSA (bin), $k = 10$	1.6:[82,50]
RR-SSA (bin), $k = 20$	1.7:[87,50]
RR-SSA (bin), $k = 30$	1.6:[70,33.33]
RR-SSA (bin), $k = 40$	1.5:[75,42]
RR-SSA (full), $k = 10$	1.6:[66,45]
RR-SSA (full), $k = 20$	1.7:[91,50]
RR-SSA (full), $k = 30$	1.7:[73,41.43]
RR-SSA (full), $k = 40$	1.9:[78,42.22]