Learning Rich Geographical Representations: Predicting Colorectal Cancer Survival in the State of Iowa

Michael T. Lash*, Yuqi Sun*, Xun Zhou†, Charles F. Lynch‡, and W. Nick Street†
*Department of Computer Science, †Department of Management Sciences, ‡Department of Epidemiology
University of Iowa
Iowa City, Iowa 52242
{michael-lash, yuqi-sun, xun-zhou, charles-lynch, nick-street}@uiowa.edu

***Abstract*—Neural networks are capable of learning rich, non-linear feature representations shown to be beneficial in many predictive tasks. In this work, we use these models to explore the use of geographical features in predicting colorectal cancer survival curves for patients in the state of Iowa, spanning the years 1989 to 2013. Specifically, we compare model performance using a newly defined metric – *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, and (c) whether a simple binary representation or richer, spectral clustering-based representation perform better. Our findings suggest that survival curves can be reasonably estimated on average, with predictive performance deviating at the five-year survival mark. We also find that geographical features improve predictive performance, and that the best performance is obtained using richer, spectral analysis-elicited features.**

I. Introduction

The rise of machine learning and corresponding advent of various deep learning methodologies in recent years hold great promise as such methods are capable of learning rich, non-linear feature representations. Such representations have been shown to be beneficial in a variety of domains, including medicine and public health. This work is concerned with methodology applied to such areas. More specifically, our focus is on exploring different representations of geographical features that can be used to predict colorectal cancer survival curves for patients in the state of Iowa.

To elaborate on such a problem, consider Figure 1, which shows colorectal cancer mortality rates, spanning the years 1989 to 2013, by zipcode tabulation area (ZCTA), for the state of Iowa. First, we wish to point out that many zipcodes have mortality rates at or above 30%, which shows the importance of accurately assessing the survival outlook of patients at the time of diagnosis, which may better inform treatment decisions [1]. Secondly, Figure 1 shows that different geographic locations experience different mortality rates. In other words, location appears to have a bearing on survival outlook.

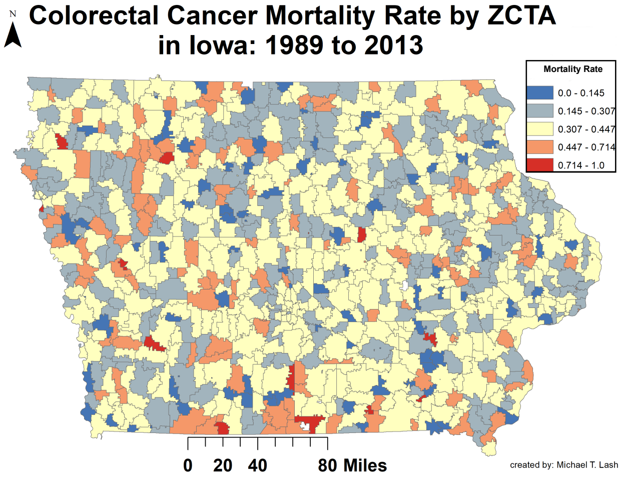


Fig. 1: Colorectal cancer mortality rate by ZCTA in the state of Iowa for the years 1989 to 2013.

Survival outlook-disparity by location is, unfortunately, not unexpected. Geographic location has been shown to have an effect on health care access, thereby affecting colorectal cancer survival outlook [2]. Moreover, environmental factors found to increase the likelihood of developing colorectal cancer tend to be spatially grouped (e.g., houses built when lead-based paint was the norm); incorporation of such factors in predictive models have been shown to provide performance improvements.

Therefore, because colorectal cancer mortality manifests itself in a spatially heterogeneous manner, a major challenge in accurately predicting colorectal cancer patient survival curves is to construct models that are spatially sensitive to patient locale: key factors affecting survival in large cities may be very different from those in rural areas. This work, therefore, explores two different ways of representing geography – termed *simple binary representation* (SBR) and *rich representation – spectral analysis* (RR-SA) – for use as features in constructing neural network-based predictive models.

The contributions of this work are enumerated as follows:

1) We investigate whether colorectal cancer patient survival curves can be reasonably predicted for patients in the state of Iowa.
2) We examine whether geographical features improve the accuracy of survival curve predictions over models trained without the use of geographic features.
3) We explore a rich representation of geographical features

978-1-5090-3050-7/17/$31.00 ©2017 IEEE

through spectral analysis (RR-SA) of the underlying adjacency graph of the ZCTAs to address the spatial heterogeneity challenge.

4) We determine whether the simple binary representation (SBR) or richer, spectral analysis representation (RR-SA) leads to more accurate survival curve predictions.
5) We propose a new metric – *area between curves* (ABC) – to assess the quality of survival curve predictions.

The remainder of this work proceeds with an outline of our methods of representing geographical features and corresponding neural network architecture associated with each such representation (Section II). Next, we discuss our dataset, which contains 46000 Iowan patients diagnosed with colorectal cancer between the years 1989 and 2013, followed by our experiments and results (Section III). Finally, we discuss related work (Section IV) prior to concluding the paper (Section V).

II. Learning Geographical Representations for Survival Curve Prediction

In this section we discuss our methodology, where we begin by outlining some preliminary notation and problem facets, followed by a discussion of Kaplan-Meier re-representation. Next, we discuss the problem of predicting individual Kaplan-Meier curves and formalize the notion of making such predictions using neural networks. The section concludes with a discussion on the different geographical representations explored in this work.

A. Preliminaries

Let $\{(\mathbf{x}^{(i)}, e^{(i)}, t^{(i)})\}_{i=1}^{n}$ 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 $t^{(i)} \in \{0, 1, \ldots, T\}$. Here, $t^{(i)}$ represents a discrete time at which the event of interest $e^{(i)}$ has occurred (i.e., $e^{(i)} = 1$) or the last discrete time instance i has been observed and the 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, however, $t^{(i)} < T$ 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. Data having the event-time representation just described are referred to as *censored data*, or more specifically *right-censored* data. An instance i is considered censored when $e^{(i)} = 0$ and $t^{(i)} < T$. Censored data, and how we handle them, are elaborated on in a subsequent subsection.

More concretely, $t \in \{1, \ldots, T\}$ might represent (as in our experiments) six-month patient follow-up periods, with $t = 0$ being the entrance of patients to the study. Entrance to the study, in this case, occurs when a patient is diagnosed with colorectal cancer. For a particular patient i, $e^{(i)} = 1$ if i dies from colorectal cancer, and $t^{(i)}$ indicates the time of this occurrence. On the other hand, an individual may move across the country, pass away from a non-colorectal cancer related complication or, for whatever reason, lose contact prior to the end of the study period. In such cases (i.e., $t^{(i)} < T$), and when patients are not known to have died from their disease, $e^{(i)} = 0$.

Each component of instance vector $\mathbf{x}^{(i)}$ represents the measurement (quantification) of a particular feature. Certain groups of these components will be referenced directly later on in this work and we therefore define notation to reference these particular groups of feature values. Let $\mathbf{z}$ contain the set of index values that index the geographical features that compose $\mathbf{x}^{(i)}$ and let $\mathbf{a}$ denote the full set of index values (i.e., $\mathbf{a} = \{1, \ldots, m\}$). These index sets will be used to reference specific components of $\mathbf{x}^{(i)}$; i.e., $\mathbf{x}_{\mathbf{z}}^{(i)}$ is the subvector of instance i containing geographical feature values, and $\mathbf{x}_{\mathbf{a}\setminus\mathbf{z}}^{(i)}$ contains non-geographical feature values.

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.
$\mathbf{g} : \mathbb{R}^m \rightarrow [0, 1]^T$	Neural network.
$\mathcal{L}(\cdot)$	An arbitrary loss function.
Smooth	Output smoothing function.
$\mathbb{Z}$	Adjacency matrix constructed from $\mathcal{M}$.
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 $\boldsymbol{\lambda}$.
$\mathbf{q}_{label}$	The result of applying kMeans clustering to $\boldsymbol{Q}_{spec}$.
Enrich	Function that assigns values in $\boldsymbol{Q}_{spec}$ to an instance.

TABLE I: Notation used throughout this work.

The notation related in this and future sections is related, for convenience, by Table I.

B. Kaplan-Meier Re-representation

With our preliminary notation defined, we return to elaborating on the censored nature of the data. As mentioned, each instance i has a corresponding event label $e^{(i)}$ and time of event occurrence $t^{(i)}$. We wish, however, to transform this tuple-like representation to one that is in the form of a *Kaplan-Meier survival curve* (KMSC) [3]. Simply put, a KMSC associates each temporal unit – in this case the values $1, \ldots, T$ – with a probability of event $e^{(i)}$ not occurring up to that particular time for instance i.

Practically speaking, this re-representation will take the form of a vector $\mathbf{y}^{(i)} \in [0, 1]^T$, where the indices $\tilde{t} \in \{1, \ldots, T\}$ denote the temporal units and the entries $\mathbf{y}_{\tilde{t}}^{(i)}$ the corresponding probabilities.

We adopt the re-representation procedure outlined in Chi et al. [4] to create $\mathbf{y}^{(i)}$, which can be expressed as

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

where $P(e_{\tilde{t}}^{(i)} = 1 | e_{\tilde{t}-1}^{(i)} = 0)$ denotes the conditional probability of event e occurring at $\tilde{t}$ given that e has not occurred at

$\tilde{t}-1$. Therefore, for patients whose outcomes are known, $\mathbf{y}^{(i)}$ contains values of 0 and 1 only, whereas a censored patient's vector becomes an estimation of survival probability at the indexical location $\tilde{t}=t^{(i)}$.

C. Predicting Individual KMSC

Ultimately, the goal of this paper is to learn a hypothesis $\mathrm{g}^* \in \mathcal{G}$, belonging to some [currently] arbitrarily defined hypothesis class $\mathcal{G}$, that most accurately predicts patient-specific KMSCs. Formally, this problem can be written as

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

where $\mathcal{L}(\cdot)$ denotes an arbitrary loss function that measures the disparity between the predicted $\mathbf{y}^{(i)}$ (in the future denoted $\hat{\mathbf{y}}^{(i)}$) and the known $\mathbf{y}^{(i)}$.

In this work, we define our hypothesis class $\mathcal{G}$ to be both shallow and deep neural networks, the specific architecture of which is elaborated on further in this section, with parameterization discussed in the experiments section. We characterize shallow architectures as having one hidden layer and deep architectures as having more than one hidden layer.

1) Output Smoothing: Neural networks are constructed in layer-wise fashion, with each layer consisting of nodes. The inputs are viewed as the first layer, followed by any number of hidden layers. The last of these hidden layers is connected to the output layer. The nature of the output layer is unique to the problem of predicting KMSCs. First, the output nodes are *ordered*. That is, we have a predicted probability for each of the $\tilde{t} = 1, \ldots, T$, where $node^{out}_{\tilde{t}}$ is *ordered* before $node^{out}_{\tilde{t}+1}$ because $\tilde{t}$ temporally comes before $\tilde{t}+1$. More importantly, however, the output elicited from these nodes should strictly decrease in temporal order. In other words, we expect $output^{(i)}_{\tilde{t}} \geq output^{(i)}_{\tilde{t}+1}$. Intuitively, even though a patient may have recovered from their disease, one would never expect the probability of survival to go up. However, because the loss function $\mathcal{L}(\cdot)$ typically produces a single value representing the loss across all nodes, the desired strictly decreasing output among temporally ordered output nodes cannot be guaranteed. Therefore, we define a smoothing procedure $\texttt{Smooth}(\mathbf{output}^{(i)})$, given by

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

which guarantees that the output elicited from the use of a trained model produces strictly decreasing outputs.

D. Geographic Feature Representation

While we are ultimately concerned with producing a g that elicits the most accurate predictions, the niche of this work is to:

1) Show whether geographic-based features improve the quality of predictions.
2) Determine whether a simple binary representation or a richer representation (defined shortly) leads to better predictions.
3) Experimentally quantify the extent of such improvements.

We outline the details of our experiments and data in the next section, where two geographic representations will be explored: a simple binary representation (SBR) and a richer representation produced via spectral analysis (RR-SA).

1) Simple Binary Representation: The simple binary representation (SBR) is a minimalist representation, involving only (a) determination of instance i's discrete geographic entity membership and (b) a binary re-representation of such membership (otherwise referred to as *one hot encoding*), producing a sparse vector with a 1 in the indexical location corresponding to the geographic entity of which i is a member, and 0s in all other locations.

To be as general as possible we assume that the current geographic features for each instance i, denoted $\mathbf{x}^{(i)}_{\mathbf{z}}$, can be used to obtain the single discrete geographic unit of which i is a member. As an example, in our experiments, we use ZCTA (zipcode tabulation area) as our discrete geographic unit.

To formalize the notion of eliciting discrete geographic unit membership, let

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

where $\Gamma(\cdot)$ is a function that transforms the geographic feature values of instance i to an ID value, denoted $x^{(i)}_b$, representing the single geography entity in a map $\mathcal{M}$ (defined shortly) that i is a member of. Depending upon the geographic information encapsulated by $\mathbf{x}^{(i)}_{\mathbf{z}}$, the function $\Gamma(\cdot)$ and map $\mathcal{M}$ may take on different forms.

In this work our geographic features are (lat,lon) coordinate pairs. Therefore, we provide a specific definition (Definition 1) outlining the map $\mathcal{M}$ that makes use of (lat,lon)-specified geography.

Definition 1. *Define $\mathcal{M}$ to be a **map**, given by*

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

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.

Map $\mathcal{M}$ is a continuous geographic region, characterized by

$$\left\{ \forall \mathtt{l} \exists \mathtt{l}' : value^q_{\mathtt{l}} = value^j_{\mathtt{l}'}, \textit{ for } \mathtt{l}, \mathtt{l}' \in \{1, \ldots, p\} \ \& \ \mathtt{l} \neq \mathtt{l}' \right\} \tag{6}$$

where $value^q_{\mathtt{l}} = value^j_{\mathtt{l}'} \triangleq (lat^q_{\mathtt{l}} = lat^j_{\mathtt{l}'}) \cap (lon^q_{\mathtt{l}} = lon^j_{\mathtt{l}'})$.

Given our definition of $\mathcal{M}, \Gamma(\cdot)$ is a function that determines whether a point, given by $\mathbf{x}^{(i)}_{\mathbf{z}}$, is on the interior of each ZCTA in $\mathcal{M}$. When the ZCTA having $\mathbf{x}^{(i)}_{\mathbf{z}}$ in the interior is found, $x^{(i)}_b$ is set equal to the ZCTA's identifier. A binarization procedure (one hot encoding), denoted $\texttt{Bin}$, is applied to $x^{(i)}_b$, thus producing a sparse vector representation.

Figure 2 illustrates the network architecture using the SBR methodology.

While we expect the addition of SBR features to elicit a hypothesis having some predictive performance improvement over a hypothesis employing only non-geographic features,

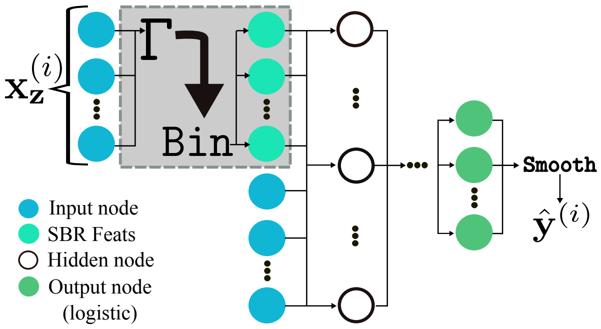


Fig. 2: SBR neural network architecture.

richer representations that better capture the continuous nature of the defined geographic region hold greater promise.

2) Learning a Rich Geographic Representation: To obtain richer geographic features, we adopt a spectral clustering-based approach (spectral analysis) to geographical feature re-representation. At a high level, this method first computes the geographic adjacency of the discrete entities that comprise $\mathcal{M}$, thus producing an adjacency matrix. Spectral analysis is then performed on this matrix. Spectral analysis involves first solving for the eigenvalues and eigenvectors of the adjacency matrix. Second, the top (i.e., largest) k eigenvalues are used to select the top k corresponding eigenvectors, forming a $p \times k$ matrix. The p rows correspond to the p geographic entities (one row corresponds to one of the p geographic entities). The k values associated with each entity are then used as predictive input features.

To express this procedure more formally, let $\mathbb{Z} = \texttt{Adj}\left(\mathcal{M}\right)$ denote the adjacency (i.e., affinity, similarity) matrix, where the l, v-th entry corresponds to the geographic adjacency relationship between the l-th and v-th discrete geographic entities, which is given by

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

where the function $\texttt{Common}(\cdot)$ evaluates whether $values_l$ and $values_v$ share a common element. In the context of the $\mathcal{M}$ described by Definition 1, $\texttt{Common}(\cdot)$ determines whether or not $values_l$ and $values_v$ have at least one coordinate pair in common.

Spectral clustering is performed by doing $\mathbf{q}_{label} = kMeans\left(\boldsymbol{Q}_{spec}\right)$, where $kMeans(\cdot)$ assigns one of k cluster labels to each of the p column elements using the k-means clustering algorithm, and where

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

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

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

Practically speaking, the column-wise elements of $\boldsymbol{Q}_{spec}$ are used as k geographical features when learning g – this is spectral *analysis* – and the labels $\mathbf{q}_{label}$ are used for visualization purposes (as in our experiments in the next section) – this is spectral *clustering*. In other words, instead of using a [necessarily] binarized form of the label assignment elicited from k-means clustering as features, we use the eigenvectors [on which clustering is performed], which preserves cluster composition.

To further differentiate spectral clustering from spectral analysis, we detail the spectral clustering procedure in Algorithm 1. Omission of the final line, highlighted in red, yields the spectral analysis procedure used to create the rich representation.

Algorithm 1 Spectral Clustering

```
Obtain adjacency matrix $\mathbb{Z}$ using (7).
Solve (9) for $\boldsymbol{Q}$ and $\boldsymbol{\lambda}$.
Obtain $\boldsymbol{Q}_{spec}$ as outlined in (8).
Apply kMeans clustering to $\boldsymbol{Q}_{spec}$ to obtain $\mathbf{q}_{label}$.
```

In other words, spectral analysis is a sub-procedure of spectral clustering, wherein the *clustering* step is omitted.

Finally, when an instance $\mathbf{x}$ is encountered, a procedure $\texttt{Enrich}(\mathbf{x}_\mathbf{z}, \mathcal{M}, \boldsymbol{Q}_{spec})$ is called that obtains the k-valued column of $\boldsymbol{Q}_{spec}$ that corresponds to the particular geographic entity that $\mathbf{x}$ belongs. $\texttt{Enrich}$ is outlined in Algorithm 2.

Algorithm 2 Enrich Geographic Features **`Enrich`**

```
Input: $\mathbf{x}_\mathbf{z}, \mathcal{M}, \boldsymbol{Q}_{spec}$
  $x_b = \Gamma(\mathbf{x}_\mathbf{z}, \mathcal{M})$ From (4).
  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 network architecture that encapsulates the spectral analysis process is shown in Figure 3[1].

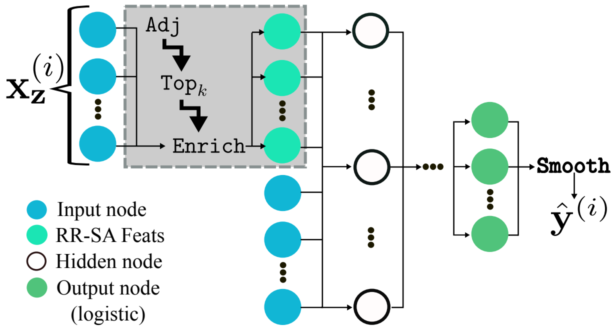


Fig. 3: RR-SA neural network architecture.

III. Predicting Colorectal Cancer Survival

In this section we begin by providing an in-depth description of the data used in our experiments, followed by an outline of the technical details of our experiments. Subsequently, we discuss experiments and results comparing average predicted survival curve against average actual survival curve by model,

[1] In our experiments $\mathbf{x}_\mathbf{z}^{(i)}$ are latitude and longitude coordinates.

as well as mean absolute error by model when the smoothing procedure is removed.

A. 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 $n = 46116$ 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 binarizing discrete features, $m = 386$ (excluding geographic features). When using SBR geographical re-representation, $m = 1364$ (386 non geographic features and $p = 978$ binarized geographic features), and $m = 386 + k$ when using the RR-SA geographic representation (where k is parameterized 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 categorized as *disease-based* and *demographic-based*. Disease-based features include tumor grade, tumor histology and tumor marker; we show a histogram of tumor grade in Figure 4. Demographic-based features include marital status, race, and age at diagnosis; we show a histogram of age at diagnosis in Figure 5. These selected features (age and tumor grade) have been shown to be indicative of not receiving timely cancer treatment [5], which we believe will help in predicting cancer survival, although analysis of such factors is beyond the scope of this work.

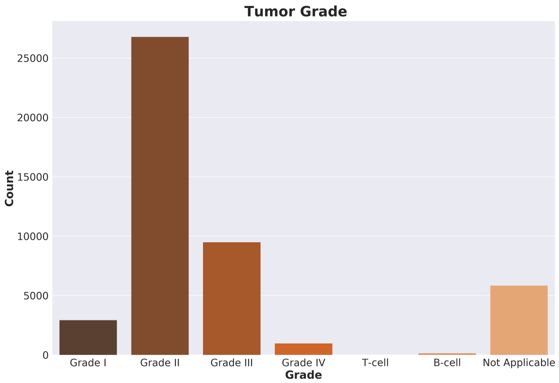


Fig. 4: Tumor grade at diagnosis for patients in the state of Iowa: Years 1989 to 2013.

B. Predictive Setting, Pamaterization 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?

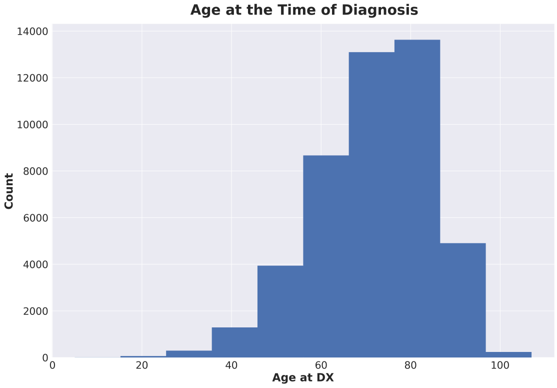


Fig. 5: Age of colorectal cancer diagnosis for patients in the state of Iowa: Years 1989 to 2013.

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?

To such an end, we propose to use 10-fold validation where, for each fold, we find a g^* for each of the following types of model:

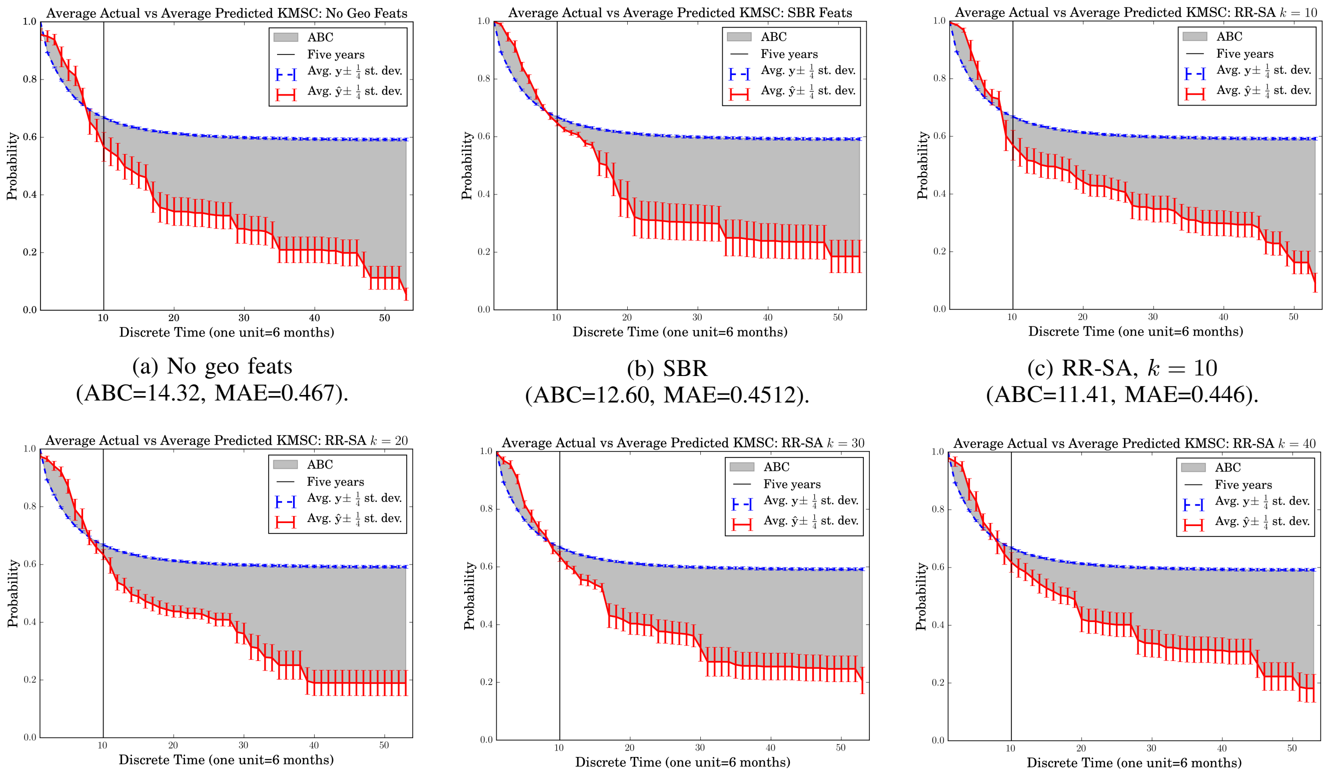
(i) A model constructed using no geographical features (No Geo).
(ii) A model constructed using SBR-derived geographical features, as outlined by Figure 2 (SBR).
(iii) Models constructed using RR-SA-derived geographical features, as outlined by Figure 3, where the values $k = 10, 20, 30, 40$ will be explored (RR-SA).

We then examine two different factors:

(a) Each model's average survival curve prediction on the test set, taken over the 10 folds, as compared to the actual average survival curve, taken over all $\mathbf{y}^{(i)}$. We devise a metric we term *area between curves* (ABC) that measures the area-wise disparity between the two curves.
(b) Each model's mean absolute error in the absence of the output smoothing procedure (described in Section 2.C.1).

1) Model Parameterization: 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 a 15% batch size. 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. Table II shows the average optimal architecture for each of the models, taken over the 10 folds.

In Table II 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 addition of rich geographic features,



(a) No geo feats (ABC=14.32, MAE=0.467).

(b) SBR (ABC=12.60, MAE=0.4512).

(c) RR-SA, $k = 10$ (ABC=11.41, MAE=0.446).

(d) RR-SA, $k = 20$ (ABC=12.31, MAE=0.453).

(e) RR-SA, $k = 30$ (ABC=11.65, MAE=0.445).

(f) RR-SA, $k = 40$ (ABC=10.77, MAE=0.442).

Fig. 6: Actual vs. Predicted.

Model	Avg 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]

TABLE II: Average optimal architecture by model over the 10 folds (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).

as defined in this work (obtained using spectral analysis), 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.

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 6. 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 value, henceforth referred to as *area between the curves* (ABC for short), as a means of comparing the predictive quality of the six different models (where lower ABC is better). We also include the mean absolute error for each model, reported as an average over the 53 outputs.

Comparing Figure 6a with Figures 6b through 6f 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. The MAE values in parenthesis support this conclusion.

Secondly, comparing Figure 6b with Figures 6c through 6f, we observe that models using richer geographical representations (RR-SA) perform better (6c - 6f) than a model trained using a simple representation (6b), again in terms of both ABC and MAE.

However, there are also RR-SA model performance differences depending on the parameterized 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 localized models, yet can also produce sparse, disjointed clusters. This point is supported by our clustering visualizations reported in Figure 8 and discussed in Section III.B.4.

In examining the different predicted survival curves we have a few observations, summarized as follows. First, we observe that predictive performance increases are mostly realized 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 realized, 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 25% (on average).
3) On average, richer geographical feature representations improve predictive performance by 15% over simpler representations.

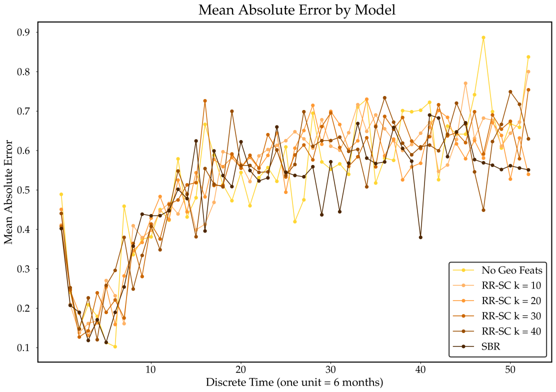


Fig. 7: MAE by model.

3) Model Errors: To further examine model performance we compare the mean absolute error of each of the models, measured at each time unit. The results comparing average error by model type are presented in Figure 7. Note that we report these error results without using the post-processing technique described in Section II.C.1 (output smoothing). We do this to provide a slightly different look at model performance over the result presented in Figure 6.

First, it is clear that all models seem to follow a similar pattern in terms of the observed error by output node, which is also found when comparing average model output in Figure 6. However, further examination reveals differences in model performance. Interestingly, SBR appears to outperform the other models at certain time point predictions toward the middle and end of the study period ($\tilde{t} \approx 30, 40$). This suggests that SBR may perform better were the smoothing method to have **not** been used. However, practically speaking, there is no circumstance in which one would want to discontinue use of such a method, but does seem to suggest that, intuitively, optimization methodology applying greater weight/emphasis to accurately learning "earlier" output nodes over "later" nodes may be beneficial.

4) Visualizing Geographic Cluster Assignment: Next, we briefly discuss the results of visualizing cluster assignment for $k = 10, 20, 30, 40$. These results can be observed in Figure 8, where each color represents a single cluster.

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 localized 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.

IV. Related Work

The topics related to and discussed throughout this work can best be categorized 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 [6]–[12] and survival-based [4], [10], [13]–[16]. The focus of this work was on survival curve predictions. Such works can be examined by method, which include Cox proportional hazards model (CPH) [13], which has been historically used to make such predictions, decision trees [14], and neural network-based models [4], [10], [15], [16], which are a more recent development. However, as Laurentiis and Ravdin [17] 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) [18]–[20]. Such methods are akin to our SBR representation and suffer from the same shortcomings. Spatially adaptive filters [21], 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 [22] and for representing geospatial features [23], [24], as in this work, and produces a rich (i.e., non-sparse) vector of features.

V. Conclusions and Future Work

In this work we explored the use of two different geographical feature representations – a simple binary representation (SBR) and a rich representation based on spectral clustering

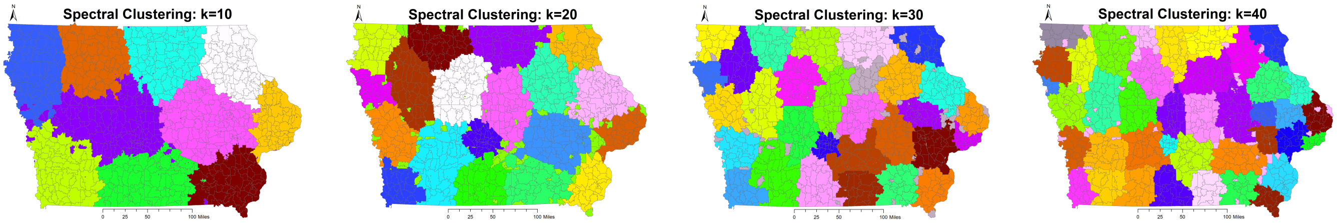

Spectral Clustering: k=10
Spectral Clustering: k=20
Spectral Clustering: k=30
Spectral Clustering: k=40
N
0 25 50 100 Miles

Fig. 8: Spectral clustering results for $k = 10, 20, 30, 40$, where color denotes cluster membership.

(which we term spectral analysis and methodologically refer to as RR-SA) – 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, and (c) RR-SA trained models outperform those trained using SBR. 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 and RR-SA geographic representations provide benefit should be undertaken.

VI. 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

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

[2] N. Wan, F. B. Zhan, B. Zou, and J. G. Wilson, "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, 2013.

[3] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[4] C.-L. Chi, W. N. Street, and W. H. Wolberg, "Application of artificial neural network-based survival analysis on two breast cancer datasets," in *AMIA Annual Symposium Proceedings*, vol. 2007. American Medical Informatics Association, 2007, p. 130.

[5] M. M. Ward, F. Ullrich, K. Matthews, G. Rushton, M. A. Goldstein, D. F. Bajorin, A. Hanley, and C. F. Lynch, "Who does not receive treatment for cancer?" *Journal of Oncology Practice*, vol. 9, no. 1, pp. 20–26, 2013.

[6] B. Khosravi, S. Pourahmad, A. Bahreini, S. Nikeghbalian, and G. Mehrdad, "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, 2015.

[7] S. Belciug, "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, 2010.

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

[9] I. K. Sandhu, M. Nair, H. Shukla, and S. Sandhu, "Artificial neural network: As emerging diagnostic tool for breast cancer," *International Journal of Pharmacy and Biological Sciences*, vol. 5, no. 3, pp. 29–41, 2015.

[10] S. Gupta, D. Kumar, and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 2, no. 2, pp. 188–195, 2011.

[11] S. Belciug and F. Gorunescu, "A hybrid neural network/genetic algorithm applied to breast cancer detection and recurrence," *Expert Systems*, vol. 30, no. 3, pp. 243–254, 2013.

[12] P. E. Puddu and A. Menotti, "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, 2012.

[13] D. R. Cox, "Regression models and life-tables," in *Breakthroughs in Statistics*. Springer, 1992, pp. 527–541.

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

[15] J. Katzman, U. Shaham, J. Bates, A. Cloninger, T. Jiang, and Y. Kluger, "Deep survival: A deep cox proportional hazards network," *arXiv preprint arXiv:1606.00931*, 2016.

[16] E. Samundeeswari and P. Saranya, "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, 2016.

[17] M. De Laurentiis and P. M. Ravdin, "A technique for using neural network analysis to perform survival analysis of censored data," *Cancer Letters*, vol. 77, no. 2-3, pp. 127–138, 1994.

[18] A. V. Khezerlou, X. Zhou, L. Li, Z. Shafiq, A. X. Liu, and F. Zhang, "A traffic flow approach to early detection of gathering events: Comprehensive results," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 6, p. 74, 2017.

[19] M. T. Lash, J. Slater, P. M. Polgreen, and A. M. Segre, "A large-scale exploration of factors affecting hand hygiene compliance using linear predictive models," in *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*, 2017 (to appear).

[20] Z. Yuan, X. Zhou, T. Yang, J. Tamerius, and R. Mantilla, "Predicting traffic accidents through heterogeneous urban data: A case study," in *6th International Workshop on Urban Computing (UrbComp 2017)*, 2017.

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

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

[23] V. Frias-Martinez and E. Frias-Martinez, "Spectral clustering for sensing urban land use using twitter activity," *Engineering Applications of Artificial Intelligence*, vol. 35, pp. 237–245, 2014.

[24] Y. van Gennip, B. Hunter, R. Ahn, P. Elliott, K. Luh, M. Halvorson, S. Reid, M. Valasik, J. Wo, G. E. Tita *et al.*, "Community detection using spectral clustering on sparse geosocial data," *SIAM Journal on Applied Mathematics*, vol. 73, no. 1, pp. 67–83, 2013.