PREDICTING THE PERCEIVED SPACIOUSNESS OF STEREOPHONIC MUSIC RECORDINGS

Andy M. Sarroff and Juan P. Bello
New York University
andy.sarroff@nyu.edu

ABSTRACT

In a stereophonic music production, music producers seek to impart impressions of one or more virtual spaces upon the recording with two channels of audio. Our goal is to map spaciousness in stereophonic music to objective signal attributes. This is accomplished by building predictive functions by exemplar-based learning. First, spaciousness of recorded stereophonic music is parameterized by three discrete dimensions of perception—the width of the source ensemble, the extent of reverberation, and the extent of immersion. A data set of 50 song excerpts is collected and annotated by humans for each dimension of spaciousness. A verbose feature set is generated on the music recordings and correlation-based feature selection is used to reduce the feature spaces. Exemplar-based support vector regression maps the feature sets to perceived spaciousness. We show that the predictive algorithms perform well on all dimensions and that perceived spaciousness can be successfully mapped to objective attributes of the audio signal.

1 INTRODUCTION

Auditory spatial impression, or the concept of type and size of an actual or simulated space [1], helps a listener form judgements about auditory events and where those events occur. In natural acoustic settings, the relative positions of sound sources to each other, the relative positions of sound sources to a listener, the listener's and sources' relative positions to the surfaces of the listening environment, and the physical composition of the structures that form and fill the listening environment are factors that contribute to spatial impression.

In a stereophonic music production, music producers seek to impart impressions of one or more virtual spaces upon the recording with two channels of audio. Spatial cues are captured, manipulated, and added in order to provide the listener with impressions of simulated acoustic spaces, whether intentionally natural or unnatural sounding. The artful handling of these cues by producers can affect enjoyability of the listening experience.

Our goal is to successfully predict the spatial impression that stereophonic recorded music imparts. A robust predictive system can empower music producers, listeners, and consumers with perceptually meaningful ways to evaluate, manipulate, and manage their music. Top-down controls for spaciousness may help music makers sculpt their sound. Casual listeners may customize their experience by using "spaciousness" controls similar to the EQ controls ubiquitous in consumer reproduction systems. By giving humans such resources, the music making and listening experience will become more flexible and interactive.

We set about our task by parameterizing the concept of spaciousness with three dimensions. A data set of stereophonic music recordings is collected and subsequently annotated for each dimension of spaciousness. We then use exemplar-based learning to build functions that map objective measurements of digital audio to the annotated music recordings.

We have structured this paper as follows: Section 2 gives some background as to how others have dealt with spaciousness and describes our approach to predicting spatial impression. In Section 3, we detail the processes of music selection and annotation. Section 4 describes the learning algorithms that are used and their parameterizations. The algorithms are tested and subsequent results are discussed in Section 5. We end with concluding remarks and suggestions for future work in Section 6.

2 BACKGROUND AND APPROACH

Our approach is summarized in Figure 1. We begin with a set of musical recordings and end with three spatial dimensions, or "target concepts"—the width of the source ensemble, the extent of reverberation, and the extent of immersion (defined in Table 1). Prediction is accomplished by mapping subjective ratings to objective measurements by machine learning.

In this paper, we focus on three relations between listener and music—the source relation (width of ensemble), the environment relation (reverberation), and the global relation (immersion). They have been selected from an amal-

SMC 2009, July 23-25, Porto, Portugal

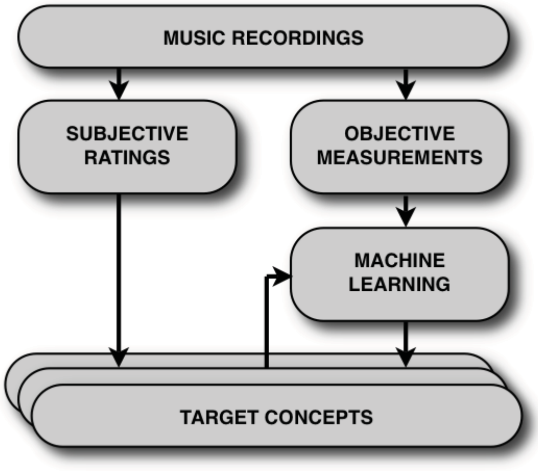

Figure 1. Framework for predicting perceived spaciousness of music recordings.

gamation of perceived attributes found in literature on natural acoustics and sound capture/reproduction. For both of these areas, there is an implicit need to rate the spatial quality of such systems. To do so, researchers must know the dimensions of spaciousness that are most salient to human listeners and to have a means of evaluating spatial quality along these dimensions.

In natural acoustics, spatial impression is divided into two primary dimensions, Apparent Source Width (ASW) and Listener Envelopment (LEV) [2, 3]. In sound capture and reproduction, the dimensionality of spaciousness is further demarcated. For example, [4] groups spatial attributes into descriptions that are related to sources, groups of sources, environments, and global scene parameters. We borrow from the literature of both fields for identifying salient spatial dimensions. ASW is defined as "the apparent auditory width of the sound field created by a performing entity" in [5] and "ensemble width" is the "overall width of a defined group of sources" in [4]. Both of these definitions connote one attribute that entails the width of the source ensemble. The extent of reverberation is directly linked to LEV in [5]. However, we use "immersion" to describe an attribute that encapsulates several kinds of envelopment, as is done in [4], and we treat immersion and reverberation independently, as is done in [6].

Once we have defined the spatial dimensions, or target concepts, we need a means of quantitatively evaluating them. In natural acoustics, objective measurements have been suggested numerous times, for example in [5] and [7]. Such measures quantify acoustic properties of physical space and relate these to perceived spaciousness. As recorded music only represents physical space virtually, measurements like these do not serve our goals. To the best of our knowledge, the perception of spatial attributes has been addressed qualitatively, but not quantitatively, in sound capture and reproduction.

- The "width of the source ensemble" of a sound is how widely spread the ensemble of sound sources appears to be.
- The "extent of reverberation" of a sound is the overall impression of how apparent the reverberant field is.
- The "extent of immersion" is how much the sound appears to surround one's head.

Table 1. Definitions of learning concepts.

It is therefor necessary for us to newly construct a set of annotated music recordings and determine a quantitative relation. The target concepts cannot be divided into a semantically meaningful finite number of categories, so we impose a bounded arbitrary continuum and build a regression model for each concept. With the exception of listener experience, perceived attributes discussed in the literature are consistently related to sound sources or their environment, rather than personal properties like gender. These are universal in nature and therefor support a model which maps spaciousness to objective measurements of the recorded signal.

3 MUSIC SELECTION AND ANNOTATION

3.1 Music Selection and Segmentation

Fifty songs were selected from an online music database [8]. The songs were equally distributed across seven genre groups: "Alt/Punk," "Classical," "Electronic-Dance," "Hip-Hop," "R&B/Soul," and "Rock/Pop." An equal propagation of chorus, verse, and bridge segments were spread across the pool. A seven second segment was excerpted from each of the songs. The duration was chosen, by informal evaluation, to be long enough to develop concrete impressions of spaciousness, yet short enough to prevent much temporal variation in spaciousness within the excerpt. None of the recordings were from commercially available songs; it is therefore unlikely that songs would be recognizable and induce bias during human annotation.

3.2 Labeling

Human subject studies were conducted online and in a laboratory. In each, subjects were required to use headphones. First, basic demographic data was collected. Participants were mostly experienced music listeners, but varied in country of residence, age, gender, profession, and other attributes of demography. Subjects were given explicit explanations and definitions of the dimensions that they were to evaluate. For each of the terms, participants were asked to listen to a non-musical mixture of sources (a room of applause). This training phase was designed to give participants time to familiarize themselves with the concepts and focus their listening on a simple stimulus. The nonmusical recordings exhibited the spatial dimensions but, to avoid pre-biasing their judgments of spaciousness, participants were not told how spacious the recordings were to be perceived.

Subjects were asked to rate, on a bipolar 5-ordered Likert scale from “Less” to “Neutral” to “More,” each of the dimensions for each song. There were a total of 98 participants providing 2,523 total ratings. Ratings were transformed from a Likert space to a numerical space by assigning the 5-ordered response categories integer values. For each song and dimension, all responses that were at least 3 standard deviations from the mean were removed as outliers. Any participant who had more than two outliers for a dimension was removed from that dimension. The responses for each dimension were standardized to zero mean and unit variance, and the mean for each dimension and song was calculated. The pairwise correlation coefficient R was calculated between ratings for the learning concepts. Width–immersion R was 0.87 and reverberation–immersion R was 0.57. Concepts width–reverberation were the least correlated, with a coefficient of 0.32, suggesting that subjects perceived differences between these dimensions unambiguously.

4 MACHINE LEARNING

A block diagram for building our objective-to-subjective mapping function is shown in Figure 2. At the beginning, we have a large feature space that objectively describes the music recordings. At the end, we have a support vector machine that needs optimization to accurately predict subjective ratings. In between, a correlation-based feature selection and subset voting scheme are used to narrow down the feature space. Then a grid search for the best parameterization of the support vector regression function is conducted. Each stage is described in detail below.

4.1 Feature Generation

A verbose set of attributes was batch-generated on the Left-Right difference signal of the data set using the MIR Toolbox [9] and two additional features. The batch-generated features include many that are widely used, like MFCCs, Spectral Centroid, and Spectral Flatness. The two additional features, which we have reported in [10], are non-standard but describe spatial characteristics of a signal. The first blindly estimates, through magnitude cancellation techniques, how widely a mixture of sources is distributed within the stereo field. The second uses the residual of a linear predictor as an indicator of how much reverberation a signal contains.

Category	**Feature**
Dynamics	RMS energy
Rhythm	Fluctuation Peak Position*, Fluctuation Peak Magnitude*, Fluctuation Spectral Centroid*, Tempo, Tempo Envelope Autocorrelation Peak Position, Tempo Envelope Autocorrelation Peak Magnitude, Attack Time, Attack Time Onset Curve Peak Position*, Attack Time Onset Peak Magnitude*, Attack Slope, Attack Slope Onset Curve Peak Position*, Attack Slope Onset Curve Peak Magnitude*
Timbre	Zero-Cross Rate, Spectral Centroid, Brightness, Spectral Spread, Spectral Skewness, Spectral Kurtosis, Roll-Off (95% threshold), Roll-Off (85% threshold), Spectral Entropy, Spectral Flatness, Roughness, Roughness Spectrum Peak Position, Roughness Spectrum Peak Magnitude, Spectral Irregularity, Irregularity Spectrum Peak Position, Irregularity Peak Magnitude, Inharmonicity, MFCCs, Δ MFCCs, $\Delta\Delta$ MFCCs, Low Energy*, Low Energy RMS, Spectral Flux
Pitch	Salient Pitch, Chromagram Peak Position, Chromagram Peak Magnitude, Chromagram Centroid, Key Clarity, Mode, Harmonic Change Detection
Spatial	Wideness Estimation*, Reverberation Estimation*
Summary Functions	Mean, Standard Deviation, Slope, Period Frequency, Period Amplitude, Period Entropy

Table 2. List of audio features, their categories, and summary functions. Features with an asterisk (*) only had their mean calculated.

For most features, the recording was frame-decomposed and feature extraction was performed on each frame. Some features, such as Fluctuation, were calculated on the entire segment. The frame-level features were summarized by their mean and standard deviation. Additionally, their periodicity was estimated by autocorrelation, and period frequency, amplitude, and entropy was calculated. The size of the final feature space extracted from the recordings was 430 dimensions. The entire set of features, which can be sub-divided into categories of Dynamics, Rhythm, Timbre, Pitch, and Spatial, is listed in Table 2.

4.2 Pre-Processing

The feature space was normalized to the range $[0, 1]$ and transformed into a principal components space. The non-principal components that accounted for the 5% least variance in the data set were discarded, and the data set was transformed back to its original symbolic attribute space.

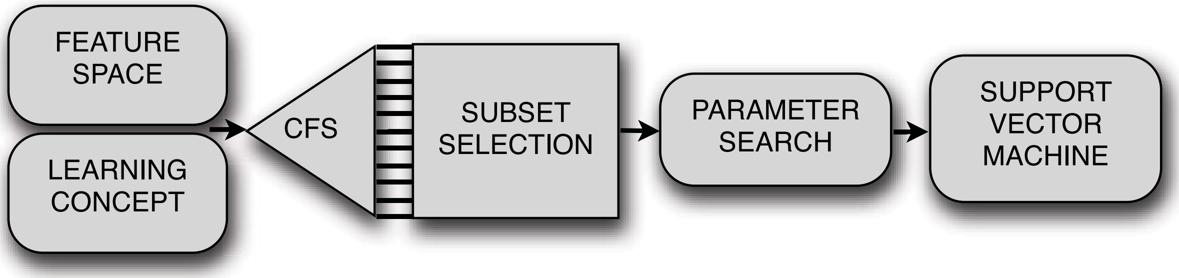

Figure 2. Block diagram for building and optimizing the mapping function.

4.3 Feature Selection

For each target concept, Correlation-Based Feature Selection (CFS) was performed with a greedy step-wise forward search heuristic. CFS chooses attributes that are well correlated to the learning target, yet exhibit low intercorrelation with each other. It has been shown to be good for filtering out irrelevant and redundant features [11].

However, supervised attribute selection can over-fit attributes to their learning concept when the same data set is used for training and testing [12]. To minimize subset selection bias, a percentile-based voting scheme with 10 × 10-fold cross-validated attribute subset selection was performed. Multiple cross-validation (CV) is a robust way of estimating the predictive power of a machine when only a small data set is available. As each fold generated a different feature set, some features were selected more often than others. For each run, features were placed in a percentile bin based upon how many times that feature had been selected. Up to 11 new data sets with monotonically increasing feature spaces were generated in this way.

Each feature space was then used to learn a non-optimized support vector regression algorithm for each dimension. The subset that performed the best for each learning concept was voted as the final subset for further system optimization and training.

4.4 Regression

For each concept, a support vector regression model was implemented with the Sequential Minimal Optimization (SMO) algorithm [13]. Support vector machines have shown to generalize well to a number of classification and regression tasks. Our support vector models employed a polynomial kernel, $K(x, y) = (< x, y > +1)^p$, chosen as the best in an informal kernel search. Support vector machines perform, to some extent, similarly well independent of kernel type if the kernel's parameters are well-chosen [14]. An exhaustive grid search for the optimal values of the support vector machine complexity (C) and its kernel exponent (p) was conducted after the optimal feature space had been selected.

5 EXPERIMENTS AND RESULTS

For each dimension of spaciousness, the best feature space was found by using Multiple CV. Then we systematically searched for the support vector parameterization that yielded the lowest error for each concept. Success was evaluated by relative absolute error (RAE), which is insensitive to scale. RAE is the sum of all the errors normalized by the sum of the errors of a baseline prediction function, Zero-R. Zero-R picks the mean value of the test fold for every instance. An error of 0% would denote perfect prediction.

Figure 3 shows the results of testing for the best feature space percentile. All predictors show two local minima: Width at the 20^{th} and 50^{th} percentiles; reverberation at the 10^{th} and 40^{th} percentiles; and immersion at the 20^{th} and 70^{th} percentiles. This indicates that there might have been more than one optimal feature subset percentile to use. We have chosen the percentile that yielded the lowest RAE for the algorithm, without testing all local minima. The steepness of the error curves between the 0 and 10^{th} percentiles shows that simply using the entire feature set without any feature selection would greatly inhibit the performance of the support vector algorithm.

The final test results are depicted in Table 3. The mean absolute error (MAE), which is dependent on scale, was no more than 0.11 for any of the predictors. The average MAE for the Zero-R predictor is shown for comparison at the bottom of the table. The predictive capability of each of the machines was well above chance, as indicated by the RAE. All predictors had a correlation coefficient R of 0.73 or higher. An R value of 0.0 would denote a complete lack of correlation between the predicted and actual values. The predictor for wideness of source ensemble performed the poorest, but still well above chance. By all measurements of accuracy, the predictor for extent of reverberation performed the best.

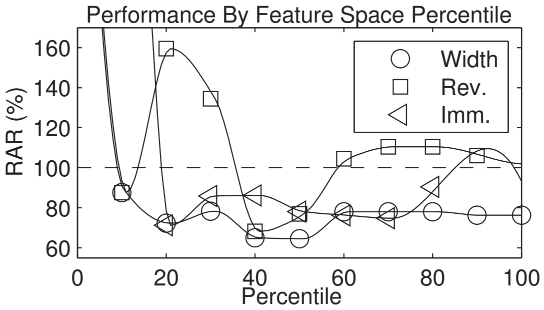

Figure 3. Performance of non-optimized machine on monotonically decreasing feature spaces.

Its coefficient of determination (R^2) indicates that the function accounted for 62% of the variance in the test set.

A summary of the final feature subset percentile used for learning each concept is shown in Table 4. While most features are probably not individually useful, the correct combination of features is. Features that were selected for more than one learning concept are shown in boldface. The width and immersion dimensions shared the most features in common; this is understandable, as these dimensions shared the highest correlation among annotations. Selected features for all three concepts were largely from the Timbre category. We find it interesting that the reverberation predictor picked three features from the Pitch category. We also note that the spatial estimators for wideness and reverberation were automatically chosen for the tasks of predicting source ensemble wideness and extent of immersion.

The error surfaces for parameterizations of each of the machines is shown in Figure 4. These surfaces show the RAE for each value in our grid search for optimal C and p values. It can be seen that the surfaces are not flat and that a globally optimal parameterization can be found for each. Yet they depict few local minima and are relatively smooth, suggesting that other parameter choices in between the grid marks would not have significantly improved results. It is worth noting that the flattest error surface, that for extent of reverberation, is also the one that performed the best, indicating robustness against parameter choices.

	Width	**Rev.**	**Imm.**
RAE(%)	62.63	67.20	64.36
MAE	0.11	0.10	0.11
R	0.73	0.79	0.76
R^2	0.53	0.62	0.58
MAE (Zero-R)	*0.19*	*0.17*	*0.18*

Table 3. The final mean absolute error (MAE), relative absolute error (RAE), correlation coefficient (R), and coefficient of determination (R^2) of the learning machines are given. The MAE for a baseline regression function, Zero-R, is given for comparison. All results are averaged from Multiple CV.

Concept (%-tile)	**Features**
Width (50 %)	**Tempo Envelope Autocorrelation Peak Magnitude Period Frequency**, **Spectral Flatness Period Amplitude**, **Wideness Estimation Mean**, **Reverb Estimation Mean**, **Δ MFCC Slope 5**, **$\Delta\Delta$ MFCC Mean 11**
Reverb. (40 %)	MFCC Mean 3, MFCC Period Entropy 3, MFCC Slope 3, $\Delta\Delta$ MFCC Period Amplitude 13, Key Clarity Slope, Chromagram Peak Magnitude Period Frequency, Harmonic Change Detection Function Period Amplitude, Spectral Flux Period Amplitude, Pitch Period Amplitude, Δ MFCC Slope 10, Δ MFCC Period Frequency 10, Δ MFCC Slope 13
Imm. (20 %)	MFCC Period Entropy 6, Spectral Centroid Period Entropy, **Tempo Envelope Autocorrelation Peak Magnitude Period Frequency**, **Spectral Flatness Period Amplitude**, Spectral Kurtosis Standard Deviation, **Wideness Estimation Mean**, **Reverb Estimation Mean**, Mode Period Entropy, Pitch Period Frequency, Δ MFCC Slope 7, **Δ MFCC Slope 5**, Δ MFCC Slope 11, Δ MFCC Mean 11, **$\Delta\Delta$MFCC Mean 11**

Table 4. Selected feature spaces after running on non-optimized machine. Features in boldface were picked for more than one learning concept.

6 CONCLUSIONS AND FUTURE WORK

We have presented a model for the automatic prediction of spaciousness in stereophonic music. We first parameterized the concept of “spaciousness” with the dimensions of source ensemble width, extent of reverberation, and extent of immersion. A verbose feature space of objective measurements was generated on a data set of human-annoted music recordings. Feature subset selection by percentile voting was used to narrow the feature space. The three target concepts were effectively learned by support vector regression with a polynomial basis function, achieving a direct mapping between signal attribute and subjective perception. Prediction for the extent of reverberation performed the best, while predictions for the wideness of the ensemble source and the extent of immersion performed slightly poorer relative to reverberation. All concept predictions exhibited RAE much better than chance.

This work is based on an assumption of independence between the learning concepts. Future work will include deeper examination of dimensional interdependency, exploration of other regressors, kernels, and feature selection algorithms, and increasing the size of our database.

The accuracies of the models suggest that objective measurements of digital audio can be successfully mapped to new dimensions of music perception. Such mappings may

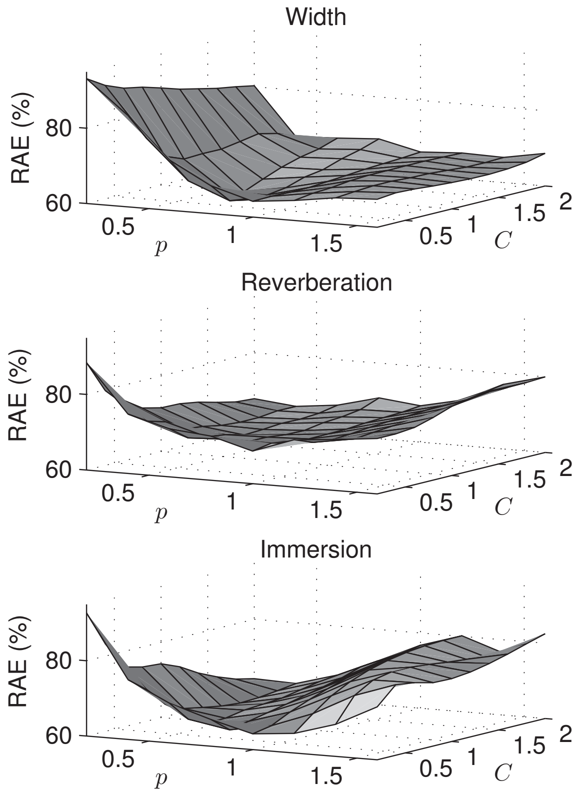

Figure 4. Relative absolute error surface for machine parameter grid search of kernel exponent p and machine complexity C.

allow music producers to have more control over their domain, including feature-driven audio synthesis and perceptually meaningful sound-sculpting. In addition, this work examines signal properties and perceptual attributes that can be tied directly to studio production of recorded music. Spaciousness is manipulated by the music engineer by applying a number of recording and signal processing techniques. By directly mapping signal attributes to the perceptual domain, music producers may gain new resources for their trade. We believe that the perceptual components of music listening that are affected by processes that occur in the production studio are a rich, yet under-exploited information stream to harvest.

7 REFERENCES

[1] J. Blauert and W. Lindemann, "Auditory spaciousness - some further psychoacoustic analyses," *Journal of the Acoustical Society of America*, vol. 80, no. 2, pp. 533–542, Aug 1986.

[2] W. Keet, "The influence of early lateral reflections on the spatial impression," in *6th International Congress on Acoustics, Tokyo, E-2-4*, 1968.

[3] M. Barron, "Late lateral energy fractions and the envelopment question in concert halls," *Applied Acoustics*, vol. 62, no. 2, pp. 185–202, 2001.

[4] F. Rumsey, "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *Journal of the Audio Engineering Society*, vol. 50, no. 9, pp. 651–666, 2002.

[5] T. Okano, L. L. Beranek, and T. Hidaka, "Relations among interaural cross-correlation coefficient (iacc(e)), lateral fraction (lfe), and apparent source width (asw) in concert halls," *Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 255–265, Jul 1998.

[6] J. Berg and F. Rumsey, "Systematic evaluation of perceived spatial quality," in *Proceedings of AES 24th International Conference on Multichannel Audio*, Banff, Alberta, Canada, 2003. [Online]. Available: http://eprints.sics.se/888/01/aes24final.doc.pdf

[7] T. Hanyu and S. Kimura, "A new objective measure for evaluation of listener envelopment focusing on the spatial balance of reflections," *APPLIED ACOUSTICS*, vol. 62, no. 2, pp. 155–184, Feb 2001.

[8] Mp3 music downloads. [Online]. Available: http://www.mp3.com/

[9] O. Lartillot, P. Toiviainen, and T. Eerola. (2008) Mirtoolbox. [Online]. Available: http://www.jyu.fi/music/coe/materials/mirtoolbox

[10] A. Sarroff and J. Bello, "Measurements of spaciousness for stereophonic music," in *125th Convention of the Audio Engineering Society, San Francisco, USA. October, 2008*, 2008.

[11] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, University of Waikato, Department of Computer Science, Hamilton, New Zealand, April 1999.

[12] A. J. Miller, *Subset Selection in Regression*. CRC Press, 2002.

[13] J. Smola, Alex and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[14] B. Scholkopf and J. Smola, Alexander, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA, USA: MIT Press, 2001.