

STAT 444 FINAL PROJECT PROPOSAL

BY ANGELO CARREON ^{1,*}, HOSEOK LEE ^{1,†}
JOY CHEN ^{1,‡} AND STEVEN SHEN ^{1,§}

¹Department of Statistics and Actuarial Science, University of Waterloo, *jaccarre@uwaterloo.ca; †h349lee@uwaterloo.ca;
‡z635chen@uwaterloo.ca; §s58shen@uwaterloo.ca

This paper contains our proposal for the STAT 444 final project. It outlines our dataset,

1. Introduction to our chosen dataset.

Our proposal primarily aims to assess the feasibility of utilizing regression techniques for interpolating and modeling housing prices. To achieve this objective, we have selected a dataset from the Journal of Statistics, which includes housing prices in Ames, Iowa, along with other pertinent features. In today's dynamic real estate market, precise and dependable house price predictions hold immense significance for homeowners, buyers, and real estate professionals alike. By delving into the potential of these advanced modeling techniques, our goal is to augment the accuracy and reliability of house price predictions, contributing to more informed decision-making in the industry.

2. Exploratory Data Analysis.

2.1. Summary Statistics.

- The dataset contains 2930 rows and 82 columns
- There are 80 explanatory variables, including 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables.

2.2. Missing/duplicate Values.

- There are four columns where missing values are extensive and need to be dropped as they are not critical to our analysis.
 - Pool.QC has only 13 observations
 - Misc.Feature has only 106 observations
 - Pool.QC has only 198 observations
 - Fence has only 572 observations
- There are no duplicate rows

2.3. Distribution of dependent variable.

- The distribution of Sale Price looks very right-skewed with most values clustered around the left tail and a significantly longer right tail, signifying extreme high values
- The sale prices range from \$12,789 to \$755,000 with a mean of \$180,796 and a standard deviation of \$79,886.69
- To achieve a more normal distribution, we will apply log transformation on the dependent variable

Keywords and phrases: We love regression.



FIGURE 1. *Histogram of Sale Price*

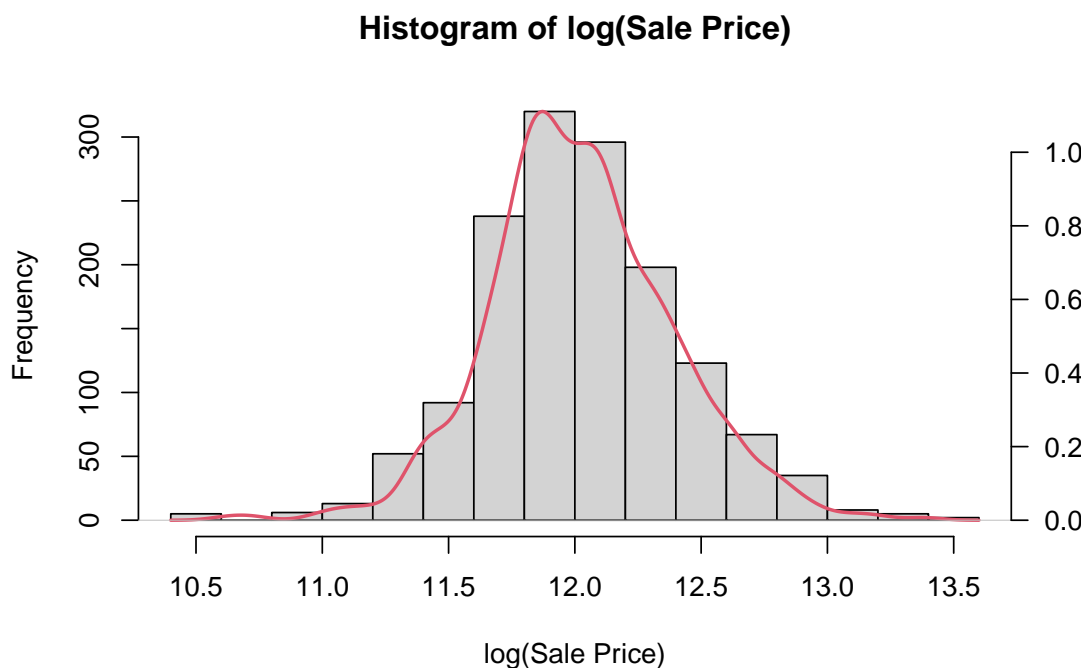


FIGURE 2. *Histogram of Log Sale Price*

2.4. *Correlations with the dependent variable.*

- Based on our correlation analysis, the top 5 numeric variables with the highest correlations with our dependent variable are:
 - Overall.Qual: Rates the overall material and finish of the house
 - Gr.Liv.Area: Above grade (ground) living area square feet
 - Garage.Cars: Size of garage in car capacity
 - Garage.Area: Size of garage in square feet
 - Total.Bsmt.SF: Total square feet of basement area

2.5. *Check if there is a high degree of correlation or linear association among independent variables.*

- Some independent variables exemplify a strong correlation, which are to be expected because they essentially provide similar information.
 - For example, the high correlation between the variable **Garage Area** (size of garage in square feet) and **Garage Cars** (size of garage in car capacity) is not surprising.
 - To address this issue of redundancy, we will select one of them and exclude the other from the analysis.

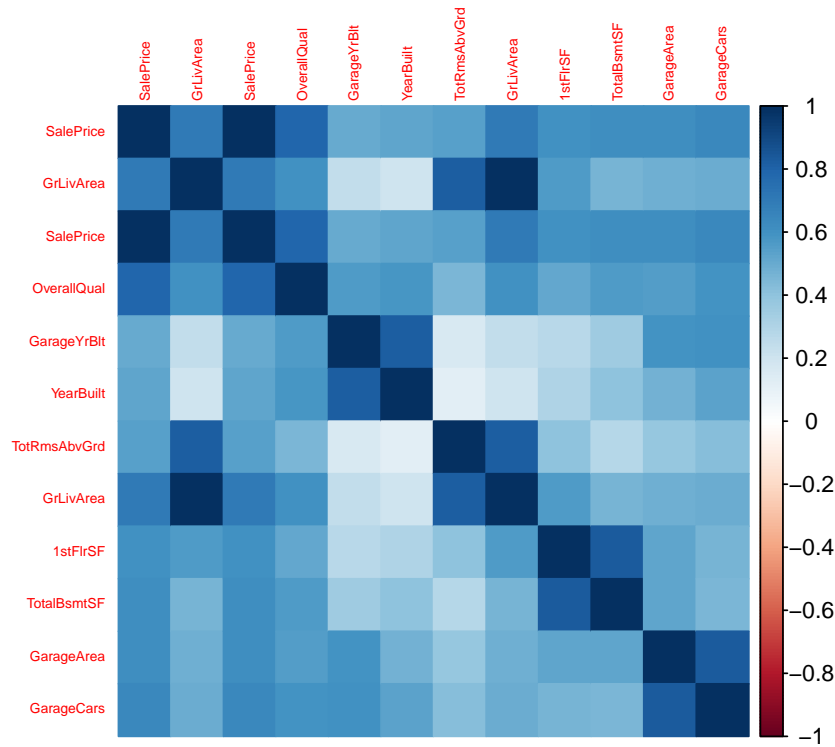
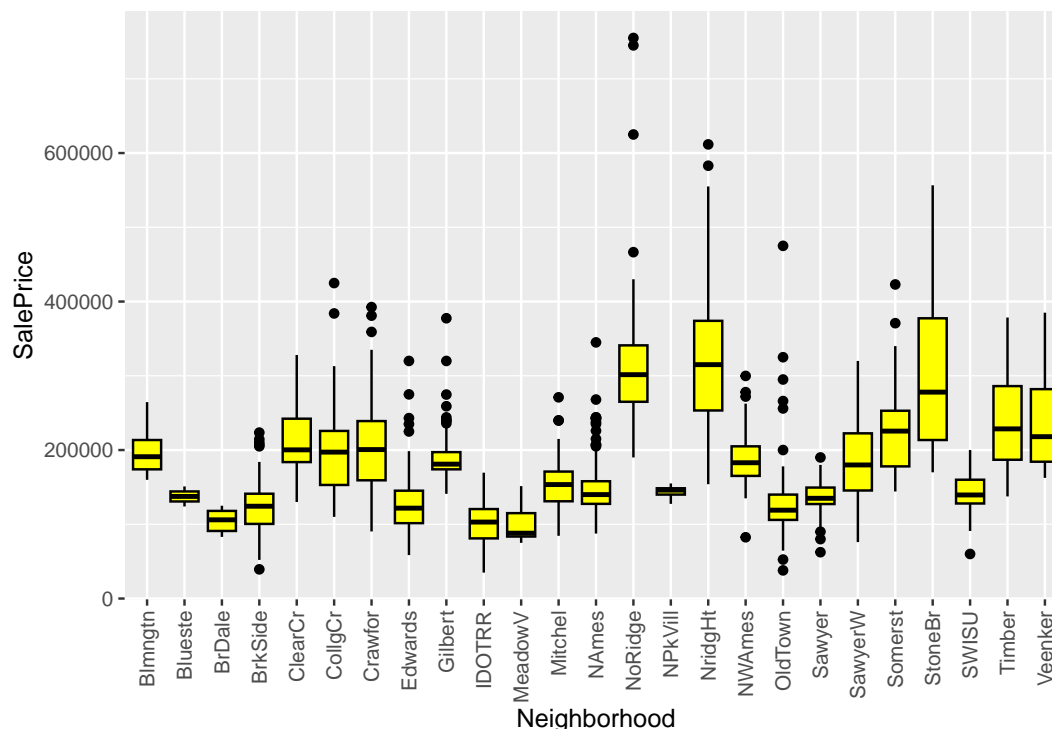


FIGURE 3. Correlation matrix of independent variables

2.6. *Other interesting but non-numeric variables.* We also assessed some of the non-numeric variables in the dataset and discovered some interesting results. * Neighborhood + There appears to be a significant difference in the average property sale price from neighborhood to neighborhood. * The dataset contains a good number of categorical variables with little variability in distribution and likely holds no predictive power + Street: Type of

road access to property + Utilities: Type of utilities available + Roof.Matl: Roof material



3. Plan.

The primary objective of this feasibility study is to determine the viability and effectiveness of employing additive models, splines, and polynomial regression for predicting house prices. Specifically, we will evaluate the following:

- **Accuracy:** Assess the predictive accuracy of additive models, splines, and polynomial regression in comparison to traditional regression models commonly used in the real estate domain.
- **Flexibility:** Analyze the ability of these techniques to capture complex relationships between house price predictors, such as square footage, location, number of bedrooms, and other relevant features.
- **Interpretability:** Evaluate the interpretability and explainability of the models, ensuring that the predictions can be easily understood and justified by stakeholders. To achieve our objective, we propose the following methodology as an outline:
- **Initial Benchmark:** We will investigate the accuracy of using linear regression with simple categorical encoding using the full set of features as an initial benchmark
- **Data Preprocessing:** Perform necessary data preprocessing tasks, including cleaning, feature engineering, and handling missing values or outliers to ensure the data quality. In an attempt to determine meaningful covariates, we are proposing to use clustering techniques such as Principal Component Analysis to remove extremely collinear covariates.
- **Model Implementation:** Develop additive models, splines, and polynomial regression models using appropriate algorithms and frameworks, such as generalized additive models (GAMs) and polynomial regression libraries in R. We will also apply penalization techniques to prevent overfitting such as forward/backward selection or LASSO/Ridge regression while scaling up the number of covariates to include interaction terms in an attempt to model any non-linear relationships

- **Model Evaluation:** Assess the performance of the models using appropriate evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), and R-squared values and compare the results with benchmark models.
- **Interpretation and Explainability:** Analyze the interpretability of the models, examining the contributions of each feature and the underlying relationships identified by the models. Utilize visualization techniques to present the results in a comprehensive and understandable manner.

4. Introduction. This template helps you to create a properly formatted L^AT_EX 2_ε manuscript. Prepare your paper in the same style as used in this sample .pdf file. Try to avoid excessive use of italics and bold face. Please do not use any L^AT_EX 2_ε or T_EX commands that affect the layout or formatting of your document (i.e., commands like `\textheight`, `\textwidth`, etc.).

5. Section headings. Here are some sub-sections:

5.1. *A sub-section.* Regular text.

5.1.1. *A sub-sub-section.* Regular text.

6. Text.

6.1. *Lists.* The following is an example of an *itemized* list, two levels deep.

- This is the first item of an itemized list. Each item in the list is marked with a “tick.” The document style determines what kind of tick mark is used.
- This is the second item of the list. It contains another list nested inside it.
 - This is the first item of an itemized list that is nested within the itemized list.
 - This is the second item of the inner list. L^AT_EX allows you to nest lists deeper than you really should.
- This is the third item of the list.

The following is an example of an *enumerated* list of one level.

- (i) This is the first item of an enumerated list.
- (ii) This is the second item of an enumerated list.

The following is an example of an *enumerated* list, two levels deep.

- 1. This is the first item of an enumerated list. Each item in the list is marked with a “tick.” The document style determines what kind of tick mark is used.
- 2. This is the second item of the list. It contains another list nested inside of it.
 - (i) This is the first item of an enumerated list that is nested within.
 - (ii) This is the second item of the inner list. L^AT_EX allows you to nest lists deeper than you really should.

This is the rest of the second item of the outer list.

- 3. This is the third item of the list.

ird item of the list. \end{longlist}

6.2. *Punctuation.* Dashes come in three sizes: a hyphen, an intra-word dash like “U-statistics” or “the time-homogeneous model”; a medium dash (also called an “en-dash”) for number ranges or between two equal entities like “1–2” or “Cauchy–Schwarz inequality”; and a punctuation dash (also called an “em-dash”) in place of a comma, semicolon, colon or parentheses—like this.

Generating an ellipsis ... with the right spacing around the periods requires a special command.

6.3. *Citation.* Simple author and year cite: Billingsley (1999). Multiple bibliography items cite: Billingsley (1999); Bourbaki (1966) or (Billingsley, 1999; Bourbaki, 1966). Author only cite: Ethier and Kurtz. Year only cite: 1956 or (1956).

7. Fonts. Please use text fonts in text mode, e.g.:

Roman

Italic

Bold

SMALL CAPS

Sans serif

Typewriter

Please use mathematical fonts in mathematical mode, e.g.:

ABCabc123

ABCabc123

ABCabc123

ABCabc123 $\alpha\beta\gamma$

\mathcal{ABC}

\mathbb{ABC}

ABCabc123

ABCabc123

$\mathfrak{ABCabc123}$

Note that `\mathcal`, `\mathbb` belongs to capital letters-only font typefaces.

8. Notes. Footnotes¹ pose no problem.²

9. Quotations. Text is displayed by indenting it from the left margin. There are short quotations

This is a short quotation. It consists of a single paragraph of text. There is no paragraph indentation.

and longer ones.

This is a longer quotation. It consists of two paragraphs of text. The beginning of each paragraph is indicated by an extra indentation.

This is the second paragraph of the quotation. It is just as dull as the first paragraph.

10. Environments.

10.1. *Examples for plain-style environments.*

AXIOM 1. *This is the body of Axiom 1.*

PROOF. This is the body of the proof of the axiom above.

□

CLAIM 2. *This is the body of Claim 2. Claim 2 is numbered after Axiom 1 because we used [axiom] in \newtheorem.*

THEOREM 10.1. *This is the body of Theorem 10.1. Theorem 10.1 numbering is dependent on section because we used [section] after \newtheorem.*

¹This is an example of a footnote.

²Note that footnote number is after punctuation.

THEOREM 10.2 (Title of the theorem). *This is the body of Theorem 10.2. Theorem 10.2 has additional title.*

LEMMA 10.3. *This is the body of Lemma 10.3. Lemma 10.3 is numbered after Theorem 10.2 because we used [theorem] in \newtheorem.*

PROOF OF LEMMA 10.3. This is the body of the proof of Lemma 10.3.

□

10.2. *Examples for remark-style environments.*

DEFINITION 10.4. This is the body of Definition 10.4. Definition 10.4 is numbered after Lemma 10.3 because we used [theorem] in \newtheorem.

EXAMPLE. This is the body of the example. Example is unnumbered because we used \newtheorem* instead of \newtheorem.

FACT. This is the body of the fact. Fact is unnumbered because we used \newtheorem* instead of \newtheorem.

11. Tables and figures. Cross-references to labeled tables: As you can see in Table 1 and also in Table 2. */%

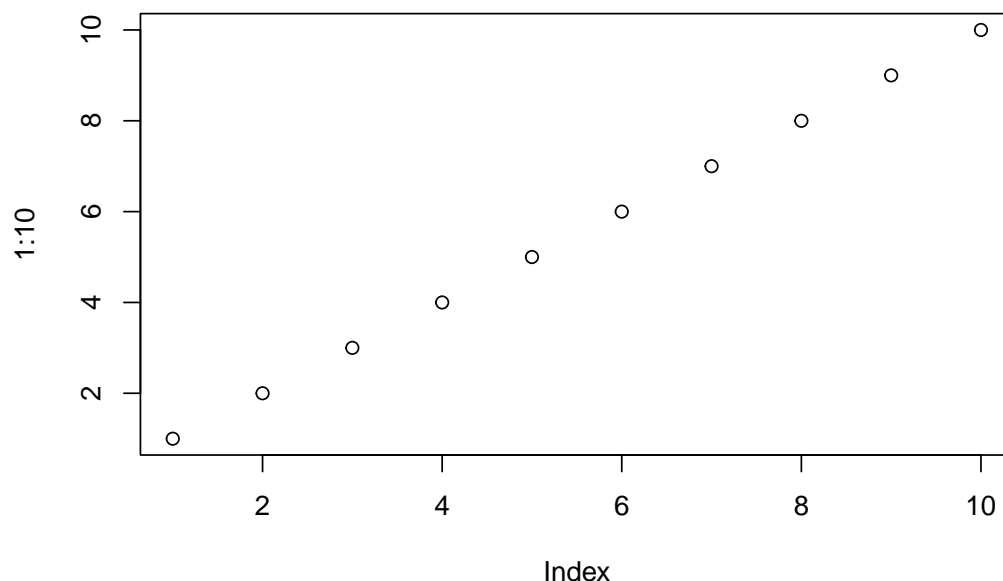


FIGURE 4. *Figure caption*

Sample of cross-reference to figure. Figure 4 shows that it is not easy to get something on paper.

TABLE 1
Table caption

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---------------------|------|-----|-------|-----|------|-------|-------|----|----|------|------|
| Mazda RX4 | 21.0 | 6 | 160.0 | 110 | 3.90 | 2.620 | 16.46 | 0 | 1 | 4 | 4 |
| Mazda RX4 Wag | 21.0 | 6 | 160.0 | 110 | 3.90 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |
| Datsun 710 | 22.8 | 4 | 108.0 | 93 | 3.85 | 2.320 | 18.61 | 1 | 1 | 4 | 1 |
| Hornet 4 Drive | 21.4 | 6 | 258.0 | 110 | 3.08 | 3.215 | 19.44 | 1 | 0 | 3 | 1 |
| Hornet Sportabout | 18.7 | 8 | 360.0 | 175 | 3.15 | 3.440 | 17.02 | 0 | 0 | 3 | 2 |
| Valiant | 18.1 | 6 | 225.0 | 105 | 2.76 | 3.460 | 20.22 | 1 | 0 | 3 | 1 |
| Duster 360 | 14.3 | 8 | 360.0 | 245 | 3.21 | 3.570 | 15.84 | 0 | 0 | 3 | 4 |
| Merc 240D | 24.4 | 4 | 146.7 | 62 | 3.69 | 3.190 | 20.00 | 1 | 0 | 4 | 2 |
| Merc 230 | 22.8 | 4 | 140.8 | 95 | 3.92 | 3.150 | 22.90 | 1 | 0 | 4 | 2 |
| Merc 280 | 19.2 | 6 | 167.6 | 123 | 3.92 | 3.440 | 18.30 | 1 | 0 | 4 | 4 |
| Merc 280C | 17.8 | 6 | 167.6 | 123 | 3.92 | 3.440 | 18.90 | 1 | 0 | 4 | 4 |
| Merc 450SE | 16.4 | 8 | 275.8 | 180 | 3.07 | 4.070 | 17.40 | 0 | 0 | 3 | 3 |
| Merc 450SL | 17.3 | 8 | 275.8 | 180 | 3.07 | 3.730 | 17.60 | 0 | 0 | 3 | 3 |
| Merc 450SLC | 15.2 | 8 | 275.8 | 180 | 3.07 | 3.780 | 18.00 | 0 | 0 | 3 | 3 |
| Cadillac Fleetwood | 10.4 | 8 | 472.0 | 205 | 2.93 | 5.250 | 17.98 | 0 | 0 | 3 | 4 |
| Lincoln Continental | 10.4 | 8 | 460.0 | 215 | 3.00 | 5.424 | 17.82 | 0 | 0 | 3 | 4 |
| Chrysler Imperial | 14.7 | 8 | 440.0 | 230 | 3.23 | 5.345 | 17.42 | 0 | 0 | 3 | 4 |
| Fiat 128 | 32.4 | 4 | 78.7 | 66 | 4.08 | 2.200 | 19.47 | 1 | 1 | 4 | 1 |
| Honda Civic | 30.4 | 4 | 75.7 | 52 | 4.93 | 1.615 | 18.52 | 1 | 1 | 4 | 2 |
| Toyota Corolla | 33.9 | 4 | 71.1 | 65 | 4.22 | 1.835 | 19.90 | 1 | 1 | 4 | 1 |
| Toyota Corona | 21.5 | 4 | 120.1 | 97 | 3.70 | 2.465 | 20.01 | 1 | 0 | 3 | 1 |
| Dodge Challenger | 15.5 | 8 | 318.0 | 150 | 2.76 | 3.520 | 16.87 | 0 | 0 | 3 | 2 |
| AMC Javelin | 15.2 | 8 | 304.0 | 150 | 3.15 | 3.435 | 17.30 | 0 | 0 | 3 | 2 |
| Camaro Z28 | 13.3 | 8 | 350.0 | 245 | 3.73 | 3.840 | 15.41 | 0 | 0 | 3 | 4 |
| Pontiac Firebird | 19.2 | 8 | 400.0 | 175 | 3.08 | 3.845 | 17.05 | 0 | 0 | 3 | 2 |
| Fiat X1-9 | 27.3 | 4 | 79.0 | 66 | 4.08 | 1.935 | 18.90 | 1 | 1 | 4 | 1 |
| Porsche 914-2 | 26.0 | 4 | 120.3 | 91 | 4.43 | 2.140 | 16.70 | 0 | 1 | 5 | 2 |
| Lotus Europa | 30.4 | 4 | 95.1 | 113 | 3.77 | 1.513 | 16.90 | 1 | 1 | 5 | 2 |
| Ford Pantera L | 15.8 | 8 | 351.0 | 264 | 4.22 | 3.170 | 14.50 | 0 | 1 | 5 | 4 |
| Ferrari Dino | 19.7 | 6 | 145.0 | 175 | 3.62 | 2.770 | 15.50 | 0 | 1 | 5 | 6 |
| Maserati Bora | 15.0 | 8 | 301.0 | 335 | 3.54 | 3.570 | 14.60 | 0 | 1 | 5 | 8 |
| Volvo 142E | 21.4 | 4 | 121.0 | 109 | 4.11 | 2.780 | 18.60 | 1 | 1 | 4 | 2 |

TABLE 2
Sample posterior estimates for each model

| Model | Parameter | Mean | Std. dev. | Quantile | | |
|---------|-----------|--------|-----------|----------|--------|-------|
| | | | | 2.5% | 50% | 97.5% |
| Model 0 | β_0 | -12.29 | 2.29 | -18.04 | -11.99 | -8.56 |
| | β_1 | 0.10 | 0.07 | -0.05 | 0.10 | 0.26 |
| | β_2 | 0.01 | 0.09 | -0.22 | 0.02 | 0.16 |
| Model 1 | β_0 | -4.58 | 3.04 | -11.00 | -4.44 | 1.06 |
| | β_1 | 0.79 | 0.21 | 0.38 | 0.78 | 1.20 |
| | β_2 | -0.28 | 0.10 | -0.48 | -0.28 | -0.07 |
| Model 2 | β_0 | -11.85 | 2.24 | -17.34 | -11.60 | -7.85 |
| | β_1 | 0.73 | 0.21 | 0.32 | 0.73 | 1.16 |
| | β_2 | -0.60 | 0.14 | -0.88 | -0.60 | -0.34 |
| | β_3 | 0.22 | 0.17 | -0.10 | 0.22 | 0.55 |

12. Equations and the like. Two equations:

$$(12.1) \quad C_s = K_M \frac{\mu/\mu_x}{1 - \mu/\mu_x}$$

and

$$(12.2) \quad G = \frac{P_{\text{opt}} - P_{\text{ref}}}{P_{\text{ref}}} 100(\%).$$

Equation arrays:

$$(12.3) \quad \frac{dS}{dt} = -\sigma X + s_F F,$$

$$(12.4) \quad \frac{dX}{dt} = \mu X,$$

$$(12.5) \quad \frac{dP}{dt} = \pi X - k_h P,$$

$$(12.6) \quad \frac{dV}{dt} = F.$$

One long equation:

$$(12.7) \quad \begin{aligned} \mu_{\text{normal}} &= \mu_x \frac{C_s}{K_x C_x + C_s} \\ &= \mu_{\text{normal}} - Y_{x/s} (1 - H(C_s)) (m_s + \pi/Y_{p/s}) \\ &= \mu_{\text{normal}}/Y_{x/s} + H(C_s) (m_s + \pi/Y_{p/s}). \end{aligned}$$

APPENDIX: TITLE

Appendices should be provided in `{appendix}` environment, before Acknowledgements.

If there is only one appendix, then please refer to it in text as ... in the [Appendix](#).

APPENDIX A: TITLE OF THE FIRST APPENDIX

If there are more than one appendix, then please refer to it as ... in [Appendix A](#), [Appendix B](#), etc.

APPENDIX B: TITLE OF THE SECOND APPENDIX

B.1. First subsection of Appendix B. Use the standard L^AT_EX commands for headings in `{appendix}`. Headings and other objects will be numbered automatically.

$$(B.1) \quad \mathcal{P} = (j_{k,1}, j_{k,2}, \dots, j_{k,m(k)}).$$

Sample of cross-reference to the formula [\(B.1\)](#) in [Appendix B](#).

Acknowledgements. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

The first author was supported by NSF Grant DMS-??-?????.

The second author was supported in part by NIH Grant ??????????.

SUPPLEMENTARY MATERIAL

Title of Supplement A

Short description of Supplement A.

Title of Supplement B

Short description of Supplement B.

REFERENCES

- BILLINGSLEY, P. (1999). *Convergence of Probability Measures*, 2nd ed. ed. John Wiley & Sons.
- BOURBAKI, N. (1966). *General Topology* **1**. Addison–Wesley, Reading, MA.
- ETHIER, S. N. and KURTZ, T. G. (1985). *Markov Processes: Characterization and Convergence*. Wiley, New York.
- PROKHOROV, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications* **1** 157–214.