A Local Weather Simulation Model and Extentions to the Transaction Model for BigPetStore

Ronald J. Nowling

December 26, 2014

Contents

1 Introduction

BigPetStore is a family of realistic example applications for the Hadoop and Spark ecosystems based on around synthetic transaction data for a fictional chain of pet stores. At the heart of BigPetStore is a data generator that implements a model for simulating customer behavior. The patterns embedded in the data as a result of the model are used as the basis for analytics examples. The incorporation of additional effects can increase the realism of the data and provide opportunities for adding more examples analyses to BigPetStore.

The data generator model aims to, where possible, incorporate effects based on *ab initio* assumptions of customer behavior. This has the advantage of making it easy to recover the resulting patterns given knowledge of the assumptions. As part of our continual effort to expand and improve the model, we want to model the effect of weather patterns on customer purchasing habits. For example, customers tend to bulk up on items before impending bad weather such as snow storms. Economic activity also tends to be reduced during cold weather. A side effect of incorporating the influence of weather in the model is the addition of regional variations in customer purchasing behavior.

Of course, to incorporate the effects of weather into the model, we first need weather data. To support simulation of arbitrary time periods, we decided to develop a simple dynamical model for generating weather data, parameterized by real, historical weather data.

2 Local Weather Model

2.1 Data Sources and File Formats

NOAA QCLCD

2.2 Temperature

To build a model, we need to a make assumptions governing the behavior of the model. By analyzing the existing data, we can infer patterns to inform our assumptions. For the analysis, we'll use the average daily temperatures from South Bend, IN between October, 2011 to September, 2014.

The average daily temperatures are plotted in Figure ??. The temperatures seem to be governed by multiple frequencies, in particular high-frequency components related to daily variations in temperature and low-frequency components related to seasonal variations. The frequency spectra of the autocorrelation of the temperatures is plotted in Figure ??. The frequency spectra is dominated by a low-frequency signal with a period of 363.3 days. The high-amplitude, low-frequency signals correspond to the seasonal change we observed in the raw data. The changes in temperature over time (derivative) (Figure 1e) appear to follow a normal distribution.

Based on the observed properties, we propose a model combining a first-order Fourier series with a period of 365 days for the low-frequency components and an Orstein-Uhlenbeck process for the noise.

$$T(t) = \frac{1}{2}a_0 + a_1 \sin\left(\frac{-2.0\pi t}{365}\right) + a_2 \cos\left(\frac{-2.0\pi t}{365}\right) + Z(t) \tag{1}$$

$$dZ_t = \theta(\mu - Z_t)dt + \sigma dW_t \tag{2}$$

Variable	*Description*
$T(t)$	Simulated temperature
t	time
a_0	Fourier coefficient
a_1	Fourier coefficient
a_2	Fourier coefficient
Z_t	Ornstein-Uhlenbeck process
θ	Damping coefficient
μ	long-term mean
σ^2	variance of the Wiener process
$dW(t)$	Weiner process

Table 1: Descriptions of variables in the model

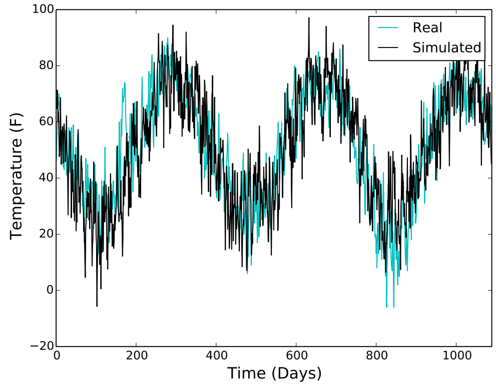
In implementing the model, we made several decisions. We determined the values of the coefficients a_0, a_1, and b_1 from a Fourier Transform of the real data. We set the long-term mean μ to 0 and determined the variance σ^2 from the distribution of the derivative values of the real temperature data.

We chose to numerically integrate the Orstein-Uhlenbeck process $Z(t)$ using the Euler-Maruyama method:

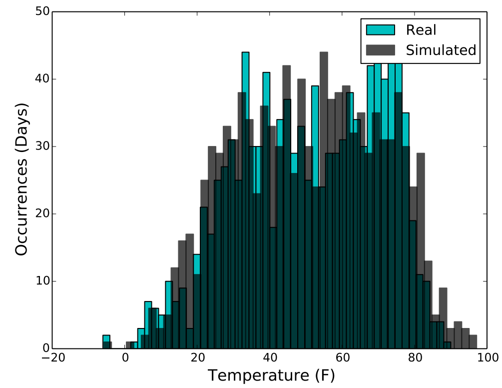
$$Z_{t+1} = -\theta Z_t \Delta t + \sigma\sqrt{\Delta t} X_{t+1} \tag{3}$$

$$Z_0 = 0 \tag{4}$$

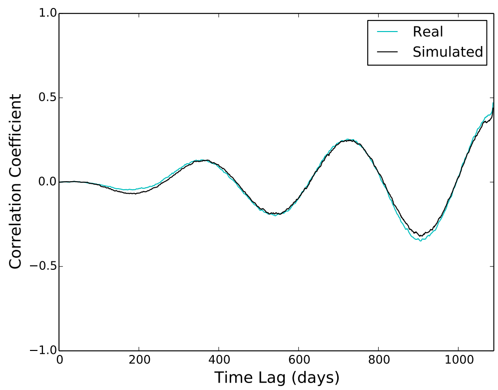
where Δt is the time step (one day) and $X_t \sim N(0, \sigma^2)$.



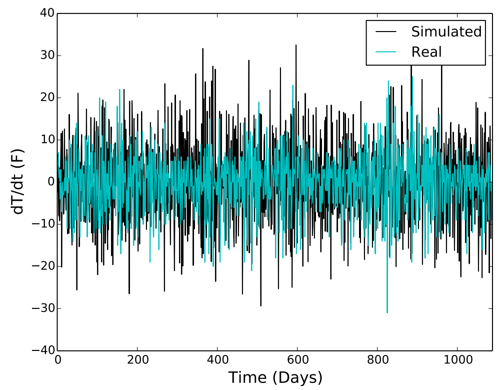
(a) Average Daily Temperature



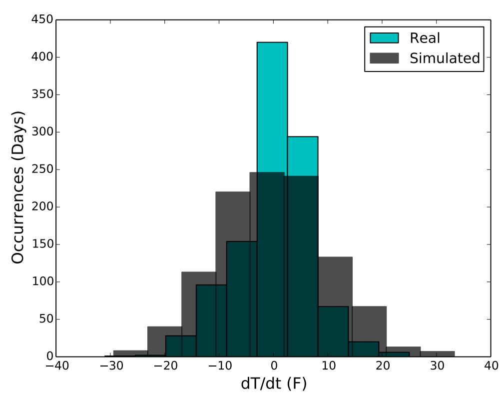
(b) Histogram of Temperatures



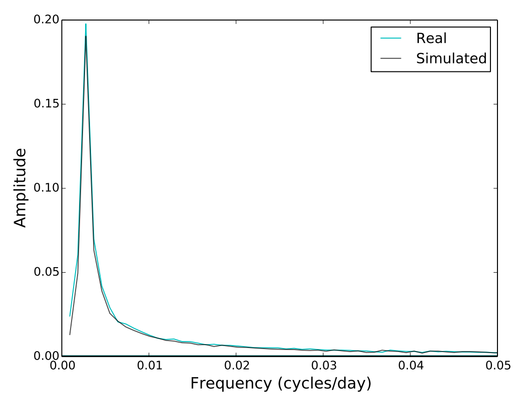
(c) Autocorrelation



(d) Derivative (dT/dt) of Temperature



(e) Histogram of Derivative (dT/dt) of Temperature



(f) Frequency Spectrum

Figure 1

2.3 Precipitation

2.3.1 Analysis of Real Data

1. Precipitation is not correlated in time. i.e., precipitation is independent day to day
2. Frequencies follow an exponential distribution
3. State (e.g., rain, snow) depends on temperature

2.3.2 Description of Model

Precipitation frequency follows an exponential distribution:

$$p(P_t|t) = \lambda \exp(-\lambda P_t) \tag{5}$$

Precipitation amounts for a given day are determined by sampling a value for P_t from $p(P_t|t)$. The precipitation is given as "water equivalent," meaning the amount of liquid water.

Snow-rain ratio is modeled by a logistic function $r(t)$ of the temperature:

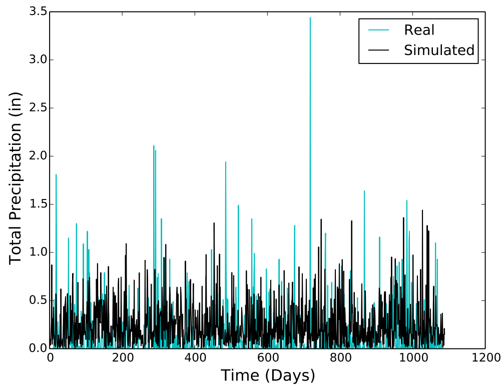
$$r(T) = \frac{1}{1 + \exp(-a(T - b))} \tag{6}$$

The amount of snowfall must be calculated from the precipitation r_t and percentage of snowfall $r(T)$. Assuming that snow has a density that is $1/10$ that of water, the amunt of snowfall is given by:

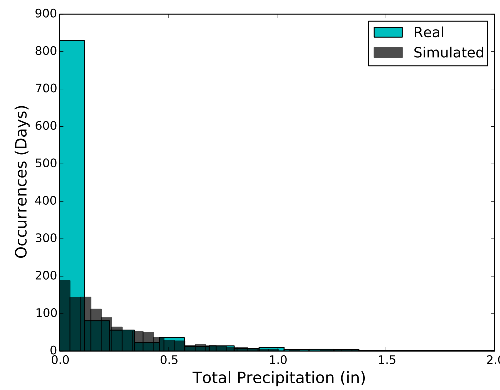
$$s(t) = 10.0\,(1 - r(T_t))\,P_t \tag{7}$$

2.3.3 Implementation of the Model

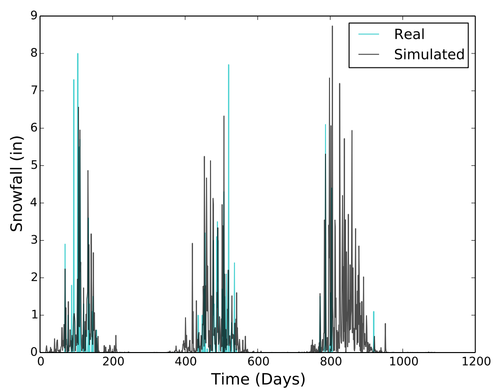
2.3.4 Evaluation of Model



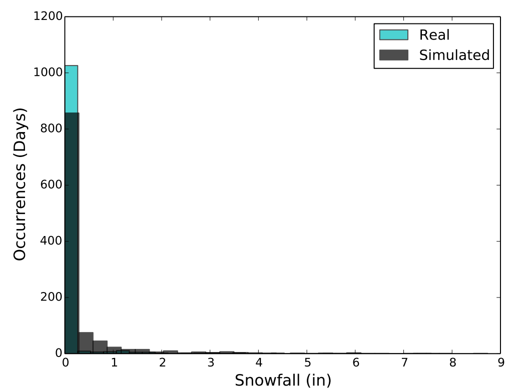
(a) Daily Precipitation



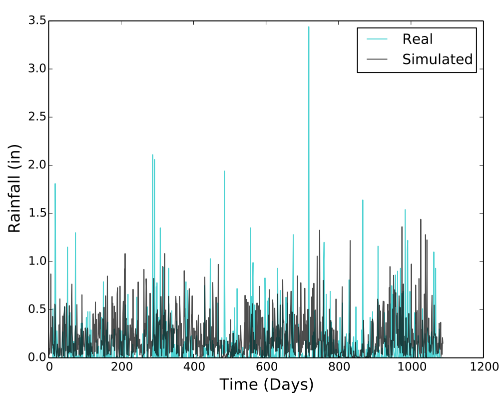
(b) Daily Precipitation Histogram



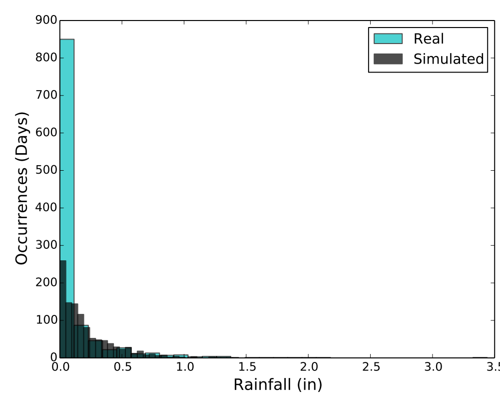
(c) Daily Snowfall



(d) Daily Snowfall Histogram

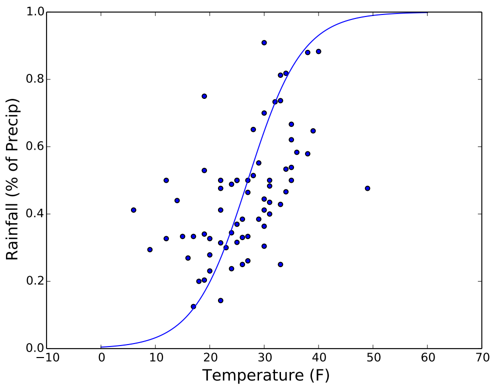


(e) Daily Rainfall

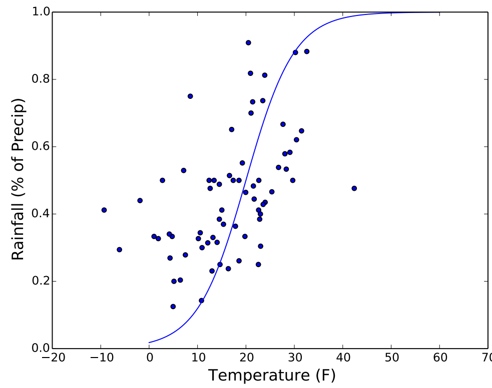


(f) Daily Rainfall Histogram

Figure 2



(a) Temperature-Precipitation Scatter



(b) Wind Chill-Precipitation Scatter

Figure 3

TODO autocorr, FFT, snowfall graph

2.4 Wind

2.4.1 Analysis of Raw Data

1. The wind process is driven by two components: a single signal in the low-frequency regime with an amplitude > the median amplitude while the other frequencies have amplitudes near the median (FFT).

2. The low-frequency signal is deterministic and sinosidal (autcorrelation).

3. The signals for the remaining frequencies are characterized by a stochastic process (autocorrelation).

4. The derivative of the wind, which is dominated by the white noise, is normally (Gaussian) distributed (derivative value histogram).

2.4.2 Description of Model

Based on the properties observed in Section **??**, we can derive a model.

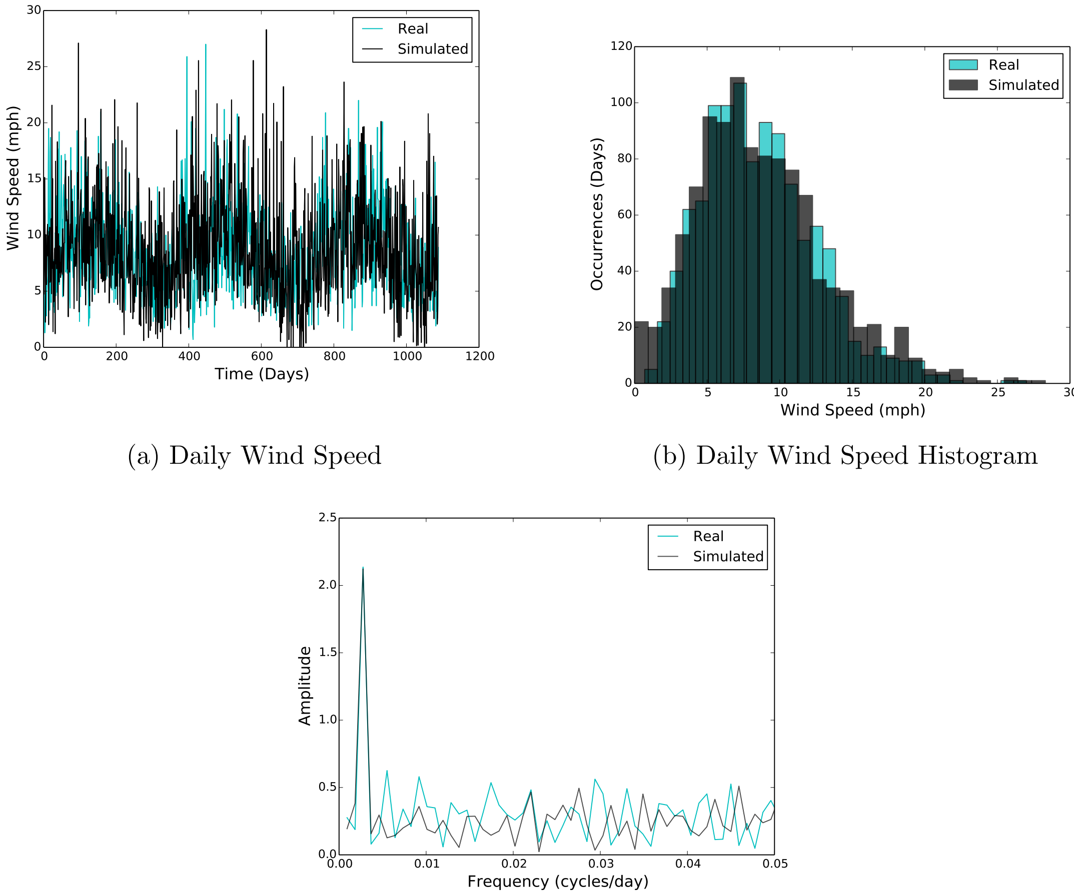
$$V(t) = \frac{1}{2}b_0 + b_1 \sin\left(\frac{-2.0\pi t}{365}\right) + b_2 \cos\left(\frac{-2.0\pi t}{365}\right) + X_t \tag{8}$$

Variable	*Description*
$V(t)$	Simulated wind speed
t	time
b_0	average wind speed
b_1	Fourier coefficient
b_2	Fourier coefficient
X_t	$X_t \sim Erlang(k, \theta)$

Table 2: Descriptions of variables in the model

2.4.3 Implementation of the Model

2.4.4 Evaluation of the Model



(a) Daily Wind Speed

(b) Daily Wind Speed Histogram

(c) Wind Speed FT

Figure 4

2.5 Review of Other Weather Models

3 Proposed Modifications to Transaction and Purchasing Models

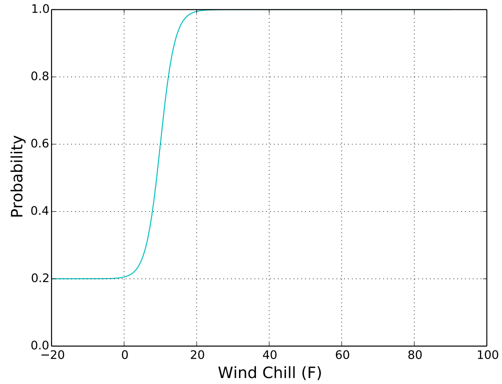
3.1 Determination of Weather Quality

Need models to quantify "quality" of weather as a value in $[0, 1]$ where 1 indicates the best possible weather and 0 indicates the worst possible weather. Weather quality could be modeled separately for temperature, snow fall, rain fall, and wind and combined using a "min" operator.

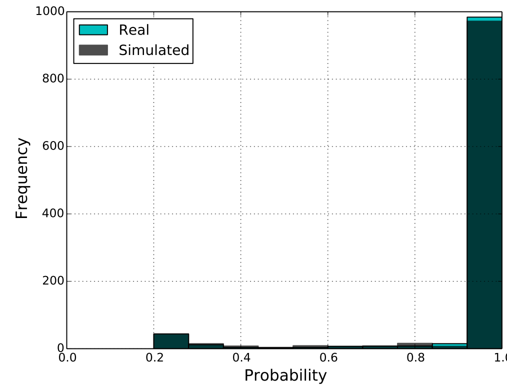
Based on the weather quality for each day, we would need to model the probability of a customer going shopping given the weather conditions.

$$q_T(T) = \frac{1-c}{1+\exp(-a(T-b))} + c \tag{9}$$

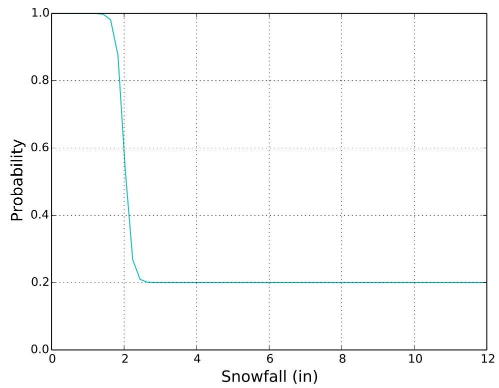
$$q_s(s) = 1 - \frac{1-c}{1+\exp(-a(s-b))} \tag{10}$$



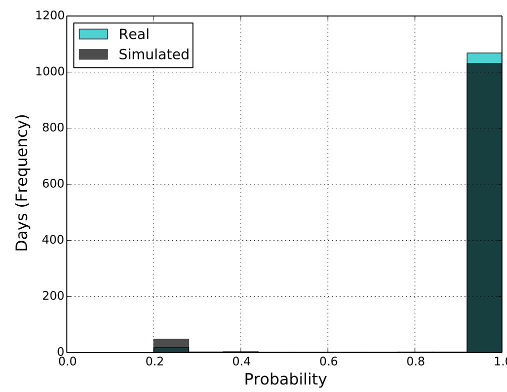
(a) Wind Chill Transaction Probabilities



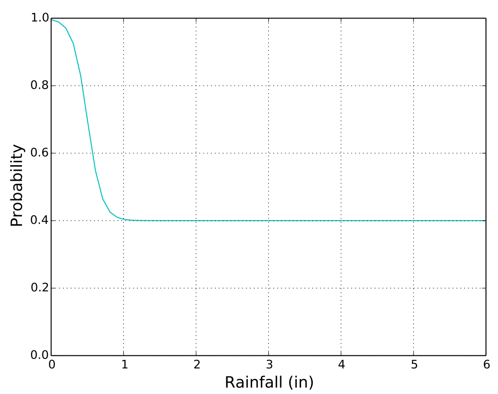
(b) Wind Chill Trans. Prob. Histogram



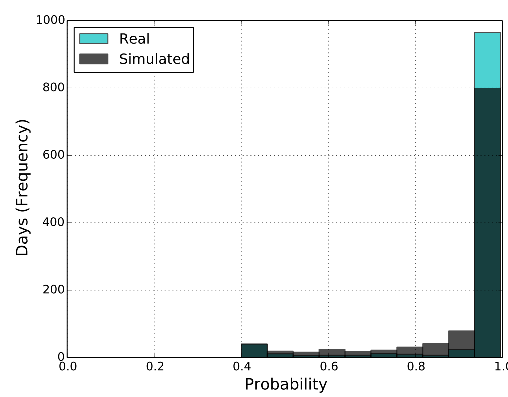
(c) Snowfall Transaction Probability



(d) Snowfall Trans. Prob. Histogram

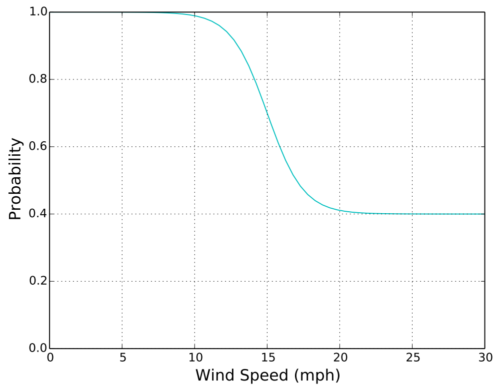


(e) Rainfall Transaction Probability

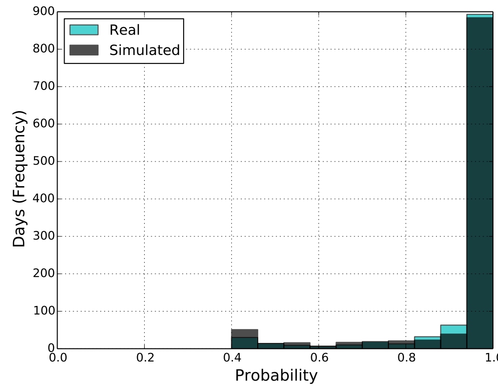


(f) Rainfall Trans. Prob. Histogram

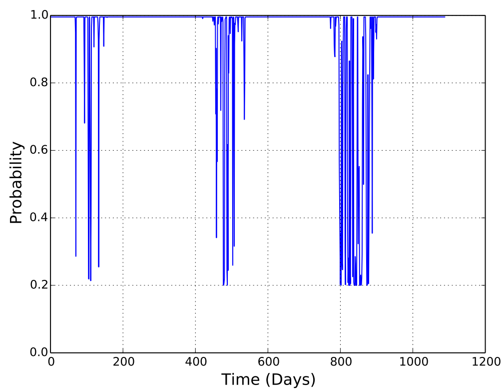
Figure 5



(a) Wind Speed Transaction Probability



(b) Wind Speed Trans. Prob. Histogram



(c) Weather Trans. Prob.

Figure 6

TODO quality vs day, total precipitation, windchill

3.2 Transaction Model

3.2.1 Review of Existing Model

3.2.2 Proposed Changes to Model

Assumptions:

1. The probability of customers going to the store on a particular day is decreased under bad weather conditions.

$$p(\text{trans}|t_i) = \frac{p_{(}\text{trans}|t_i, \text{weather})\, p_{\text{exhaustion offset}}(\text{trans}|t_i)\, p_{\text{arrow of time}}(\text{trans}|t_i \geq t_{i-1})}{Z} \tag{11}$$

3.2.3 Evaluation of Changes

3.3 Purchasing Model

3.3.1 Review of Existing Model

3.3.2 Proposed Changes to Model

Assumptions:

1. Customers will buy larger quantities if they anticipate bad weather coming

Ideas:

1. Add additional term to category weight functions based on amount of bad weather in the next N days.
2. Increase preferences for larger sizes

3.3.3 Evaluation of Changes

4 Conclusion