Tulip Germination Analysis

Jeremy Meyer

Brigham Young University
Department of Statistics

April 2019

Schedule

1. Introduction
   - background
   - goals
   - The Data
2. The Model/Algorithm
   - Proposed Model/Algorithim
   - Addressing Data concerns
3. Model Justification / Performance
   - Variable justification
   - Assumptions
   - Fit/Prediction
4. Results
   - Model Coefficients
   - Research Questions
5. Conclusion

Introduction

Background


[1]

- Holland: "Flower shop of the World"
- Tulip festivals are a tourist attraction
- 9 Million bulbs $\rightarrow$ 25% of agricultural exports

Tulip Farming Concerns

- Planted in the Fall since they require chilling time for growth
- Climate change threatens tulip economy
  - More precipitation / Flooding of low lying areas problematic for growth
  - Temps expected to rise twice the global average
- Rising temperatures may not allow for optimal chilling time

The problem

- Typical chilling time is 10 weeks
- Researchers are interested what conditions are ideal for tulip species amidst climate changes
- Specifically, we will look at germination under various chilling periods for 12 species

The Data

- 210 bulbs of 12 different species of tulip (2510 total).
- Seeds collected across a period of 5 years (2013-2017).
- 30 from each species were given a chilling time (0, 2, ..., 12 weeks).
- Response was if they bloomed.

Goals of the study

We will address:

1. Is the probability of germination for each chilling time the same across all populations? If so, which ones are similar/different?
2. Is there an “ideal” chilling time? Does this ideal chilling time vary by population?
3. What effect will a decrease in weeks of winter/chilling time have for tulips? Is it the same for each population?

Dates Collected

Year Collected by Population

	1	2	3	4	5	6	7	8	9	10	11	12
2013	0	0	0	0	0	0	0	210	0	0	0	0
2014	0	210	0	0	0	0	0	0	0	0	0	0
2015	210	0	0	0	0	0	0	0	0	0	0	0
2016	0	0	0	0	0	210	210	0	210	210	210	0
2017	0	0	210	210	210	0	0	0	0	0	0	210

Year Collected by Day (May 18 - Sep 24)

	138	161	162	164	191	198	202	203	204	230	267
2013	0	0	0	0	0	0	0	0	0	210	0
2014	0	0	0	0	210	0	0	0	0	0	0
2015	210	0	0	0	0	0	0	0	0	0	0
2016	0	0	0	0	0	210	210	210	420	0	0
2017	0	210	210	210	0	0	0	0	0	0	210

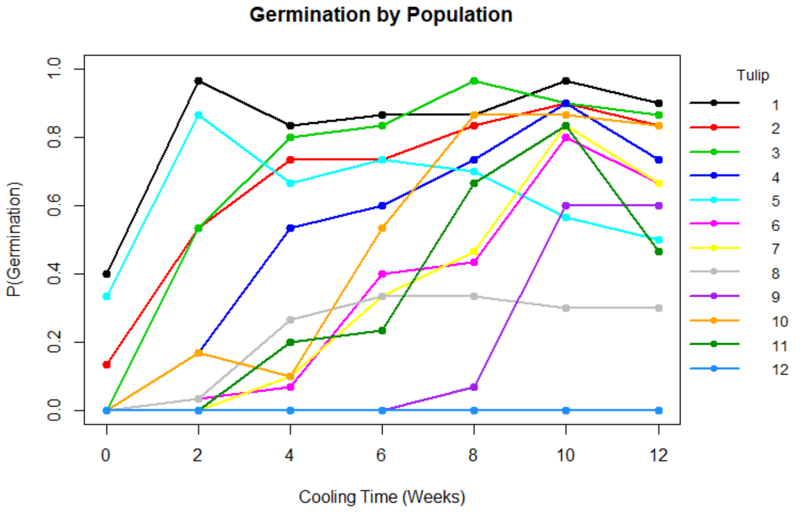
- Population groups collected all at same time.
- If these affect the response, they are confounded with the populations.

Contingency Table (%)

Bulb germination percentage across chilling periods

Species	0 Wk	2 Wks	4 Wks	6 Wks	8 Wks	10 Wks	12 Wks
#1	40.0	**96.7**	83.3	86.7	86.7	**96.7**	90.0
#2	13.3	53.3	73.3	73.3	83.3	**90.0**	83.3
#3	0.0	53.3	80.0	83.3	**96.7**	90.0	86.7
#4	0.0	16.7	53.3	60.0	73.3	**90.0**	73.3
#5	33.3	**86.7**	66.7	73.3	70.0	56.7	50.0
#6	0.0	3.3	6.7	40.0	43.3	**80.0**	66.7
#7	0.0	0.0	10.0	33.3	46.7	**83.3**	66.7
#8	0.0	3.3	26.7	**33.3**	**33.3**	30.0	30.0
#9	0.0	0.0	0.0	0.0	6.7	**60.0**	**60.0**
#10	0.0	16.7	10.0	53.3	**86.7**	**86.7**	83.3
#11	0.0	0.0	20.0	23.3	66.7	**83.3**	46.7
#12	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tulip Data


Germination by Population
P(Germination)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
Cooling Time (Weeks)
Tulip
1
2
3
4
5
6
7
8
9
10
11
12

Data problems?

- Response is binary, two predictors:
  1. Cooling Time (Numeric)
  2. Population (Factor)
- Cooling time effect on growth is inconsistent across tulip populations $\rightarrow$ Interactions
- Some tulip populations eventually decrease for high cooling times. $\rightarrow$ Non-monotonic relationships

The Model/Algorithm

Model Statement

Model

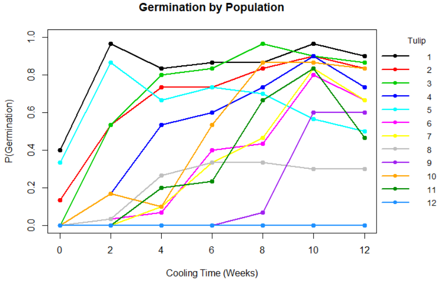
$$Y_i \stackrel{ind}{\sim} \text{Bernoulli}(p_i), \quad \boldsymbol{x_i'}\beta = \log\left(\frac{p_i}{1-p_i}\right)$$

Y_i: the response for the ith tulip
p_i: the probability of the ith tulip blooming
$\boldsymbol{x_i'}$: vector of covariates (including basis function expansions) for the ith tulip
β: coefficients of the covariates, the effect of the covariate on the log-odds.

Why Logistic regression?

- It allows us to quantify uncertainty!
  - Inference-type research questions
- Captures categorical outcomes
- Predicted probabilities for each tulip (research questions)
- Can address data issues by **interactions** and **basis function expansions**

Addressing Non-Monotone Relationships



Problem: Probability of germination eventually goes down for some tulips

- Log odds linear $\rightarrow$ "S curve"
- Basis function expansion on cooling time variable.
- We can use **Natural Cubic Splines** to capture non-monotonicity.
- Natural Cubic splines for stable end-tail behavior.

Natural Cubic Splines

Basis function expansion of Cubic Splines

For one tulip:

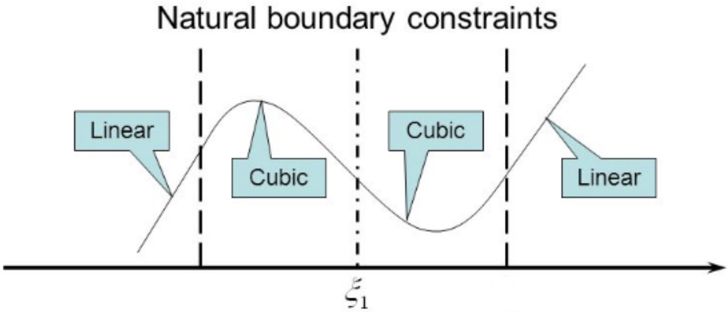
$$\boldsymbol{x_i'}\boldsymbol{\beta} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \sum_{k=1}^{K} (x_i - \xi_k)_+^3 \beta_{k+3}$$

$$(x_i - \xi_k)_+^3 = (x_i - \xi_k)^3 I(x_i > \xi_k)$$

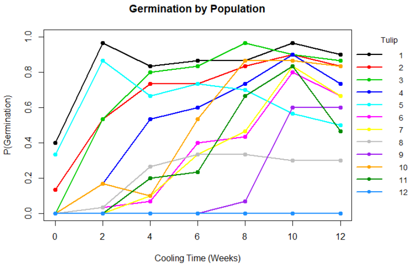
x_i cooling time for the i^{th} tulip
β_i Are the basis function coefficients
K is the number of breaking points or knots
ξ_k represents the value of the k^{th} knot in the data.
I( ) is an indicator function (0 or 1).

- **Natural splines** have extra linear constraints at endpoints.
  - Extrapolation after 12 weeks may be useful
  - 2 knots would need $(1 + 3) + 2 - 2 = 4$ β coefficients.

Natural Splines Concept



Interactions



Problem: Tulip growth is not the same across cooling times between populations.

- Intercept for each tulip population
- Include a population/cooling effect interaction term
- Fits a different natural spline for each tulip population
- Allows populations to have different germination rates over time.

Model Continued

$$\boldsymbol{x_i'}\boldsymbol{\beta} = \beta_1(\text{Tulip-1})_i + \beta_2(\text{Tulip-2})_i + ... + \beta_{12}(\text{Tulip-12})_i + \text{Tulip1:ns(Cooling Time, K)}_i$$
$$+ \text{Tulip2:ns(Cooling Time, K)}_i + ... + \text{Tulip12:ns(Cooling Time, K)}_i$$

Notes

Tulip-*n*: Indicator if i^{th} tulip is in the n^{th} population
- No overall intercept, just one for each population ($\beta_1, ..., \beta_{12}$)

Cooling Time: Numeric, number of weeks seed was cooled
ns(..., K): Cubic Natural spline with K knots
Tulip-*n*:ns(Cooling Time, K): Interaction between the populations and cooling time.
- One spline per population
- This is zero unless it matches the i^{th} tulip's population
- Total: 12(K+1) additional β coefficients.

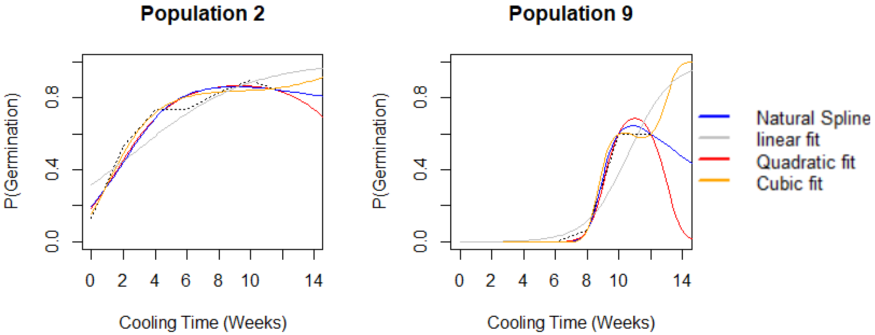
Model Justification / Performance

Variable Justification

- We did not include Year/Day because they were confounded with the populations
- We used splines, we could have also used a quadratic polynomial.
- We must also choose the number of knots/knot location.

Why not just polynomial fit?

- Can't fit an equilibrium after 12 weeks (model flexibility)
- Erratic behavior beyond range of data.



Number of Knots: Information Criterion

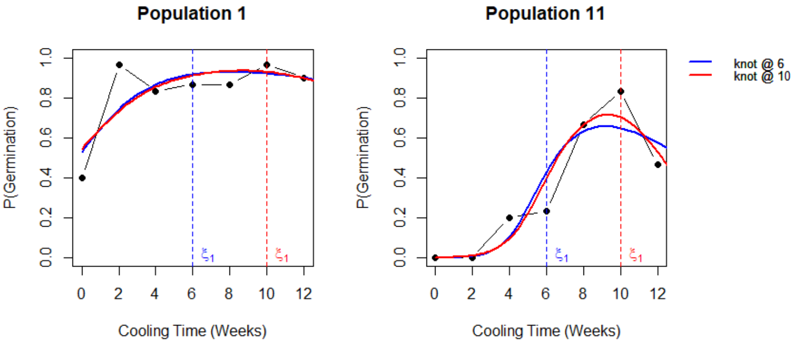
- 1 knot = 12 more parameters
- Cautious of overreacting to noise: BIC
- Lower standard errors for research questions

# Knots	df	BIC
No Knots	24	2328.922
1 Knot	**36**	**2295.448**
2 Knots	48	2366.646
3 Knots	60	2417.062
4 Knots	72	2490.950

Knots were placed according to data quantiles

Knot location?

- Defaults to median of chilling time (6 weeks).
- Knot at week 10 (BIC 2294.6) was best able to capture the peak.



Differences are relatively small, but we'll go with 1 knot @ week 10

Model Assumptions

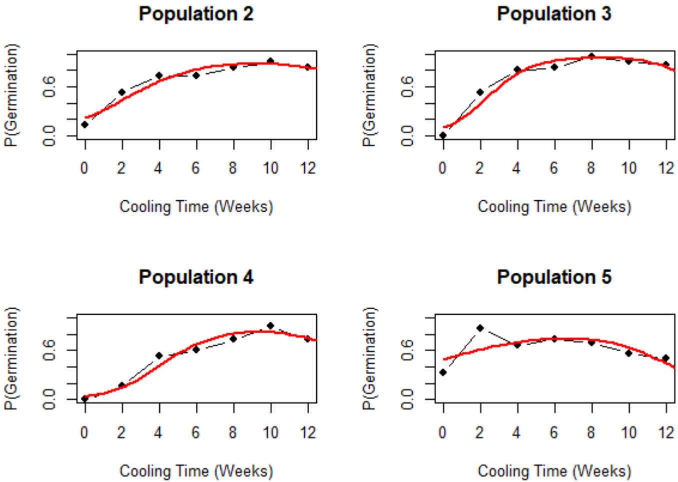
1. Independence
2. Monotonicity? (Kind of)
3. Multicollinearity

Independence

- Each Tulip's germination is independent of the others after taking into account explanatory variables.
- Reasonable assumption, but could check how the seeds were obtained
- Unwanted lab conditions

Monotonicity?

- Relationships are not monotonic!
- Instead, make sure fitted splines (red) reasonably fit with the data (black).



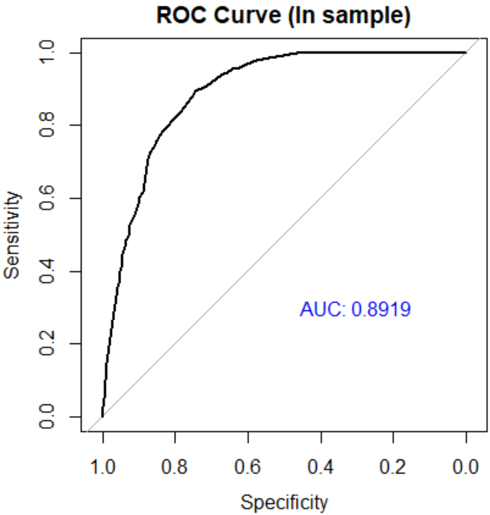
Multicolinearity

- GVIFs were 4.24 (population) and 2.06 (Cooling Time:Population).
- No surprise because of the interaction.
- However, we are fitting a separate spline per tulip population, which is equivalent to fitting 12 independent spline models with an intercept.
  - Estimates and Standard Errors are identical!
- In that case, only 1 predictor: Chilling Time.
- Thus, multicolinearity is not a problem

Model Fit

- We need Classification accuracy
- ROC (Receiver Operating Characteristic) Curve
  - Uses many different cutoff values
  - Plots sensitivity (true positive rate) against specificity (true negative rate)
- AUC (Area Under the Curve) summarizes the ROC curve

Model fit ROC(s)



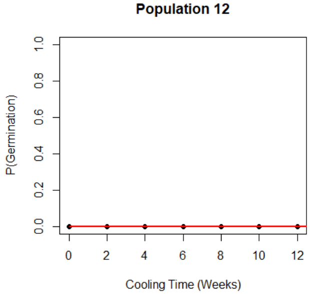
- AUC = 1: Perfect Classification
- AUC = 0.5: Coin Flipping rate
- The model does pretty well!

Choosing a Cutoff

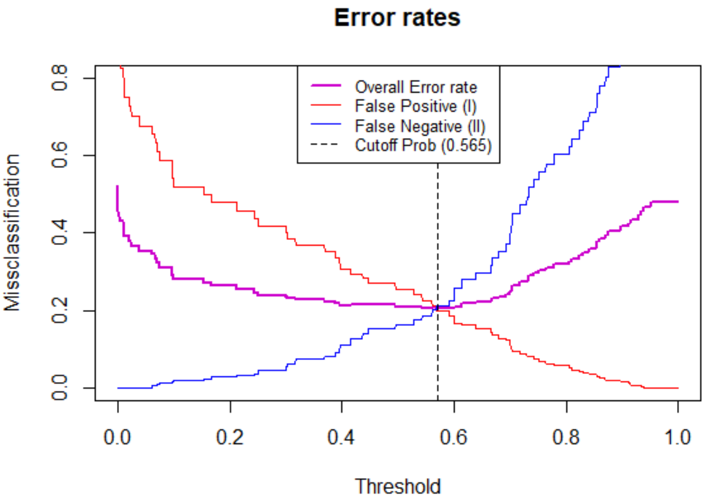
- In order to classify, we must choose a cutoff value
- “Best" cutoff depends on the goals of the analysis
- For tulip classification, unknown which type of error is worse
- We will seek to equalize type I and type II errors.

Problem: Population 12

- None germinated in the data
- Every predicted probability is $2.35 \times 10^{-8} \rightarrow$ always classified as no growth
- These will all appear to be classified correctly (Inflated Specificity)
    - These will be thrown out to determine the appropriate cutoff.



Optimal Cutoff Probability



In-sample Confusion Matrix

Table: Confusion Matrix, cutoff = 0.565

	Predicted Germination	Predicted No Growth	Total
True Germination	883	222	1105
True No Growth	257	948 (1175)	1205 (1415)

- Sensitivity = 883/(883+222) = 0.799
- Specificity = 948/(948+257) = 0.787 (w/pop 12: 0.818)
- Overall Accuracy: (883+948)/(883+948+222+257) = 79.3%
- With population 12, overall accuracy is **81.0%**

How well does this predict? (Cross Validation)

Separate splines for each population $\rightarrow$ **stratified CV**

- Fair comparisons of accuracies between groups

1. Randomly sample 70% from all 12 populations for a training set. (Stratified)
2. Compute AUC and various accuracy rates on the left out 30% (testing set)
3. Repeat 1-3 for 1000 iterations. Average across all AUC/accuracy rates.

Predictive accuracy

- Average AUC $\rightarrow$ **0.8824**

Table: Average Confusion Matrix Percentage, cutoff = 0.565

	Predicted Germination	Predicted No Growth
Germination	78.6%	21.3%
No Growth	10.9%	81.0%

- Average Sensitivity = 78.6%
- Average Specificity = 81.0%
- Average Overall Accuracy = **80.0%**

Test CV accuracy across species

- We can also check how the model does across species:

Tulip	Accuracy (%)
1	83.4
2	76.5
3	83.6
4	75.2
5	64.8
6	76.7
7	78.4
8	77.5
9	83.5
10	81.7
11	78.2
12	100.0

Notes:

- Species 5 really struggles the most with prediction
- Species 12 is never predicted to germinate
- If Tulip 12 is taken out, we only have **78.1%** overall accuracy.

Results

Interpretation of Coefficients

- 36 Total β coefficients!
  - Two types: Intercepts and spline coefficients
- 12 intercepts -> One intercept per population
- Intercept coefficients represent germination log-odds probability after no cooling time (0 Weeks).
- We expect about 95% of the confidence intervals to contain the true proportion of germination at time 0.

Intepretation of Intercepts

Assuming cooling time is zero, the intercept for the i^{th} population, β_i

$$\log(\frac{p}{1-p}) = \boldsymbol{\beta_i} \quad \rightarrow \quad p = \frac{1}{1+e^{-\beta_i}} \tag{1}$$

Estimated Log-odds probability at Time = 0

Tulip	$\hat{\beta}_i$	2.5%	97.5%
1	0.171	-0.463	0.818
2	-1.309	-2.045	-0.638
3	-2.224	-3.174	-1.402
4	-3.207	-4.429	-2.193
5	-0.016	-0.620	0.589
6	-5.688	-8.316	-3.745
7	-6.666	-9.830	-4.335
8	-3.714	-5.307	-2.479
9	-37.788	-64.282	-17.465
10	-4.484	-6.298	-3.060
11	-7.019	-10.001	-4.727
12	-17.566	-191.765	12.270

Estimated probability at Chilling Time = 0

Tulip	Estimate	2.5%	97.5%
1	0.543	0.386	0.694
2	0.213	0.115	0.346
3	0.098	0.040	0.198
4	0.039	0.012	0.100
5	0.496	0.350	0.643
6	0.003	0.000	0.023
7	0.001	0.000	0.013
8	0.024	0.005	0.077
9	0.000	0.000	0.000
10	0.011	0.002	0.045
11	0.001	0.000	0.009
12	0.000	0.000	1.000

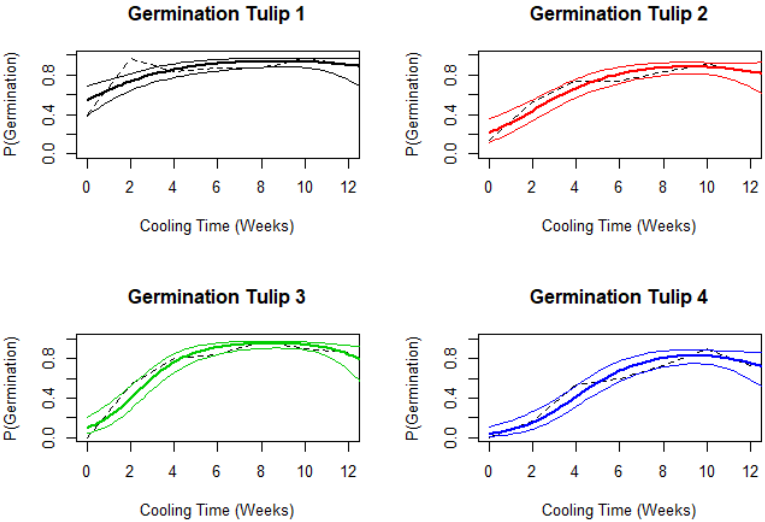
Interpretation of Spline Coefficients

- Ugly basis function expansions → difficult to interpret
- There are 24 total, 2 spline coefficients for each population

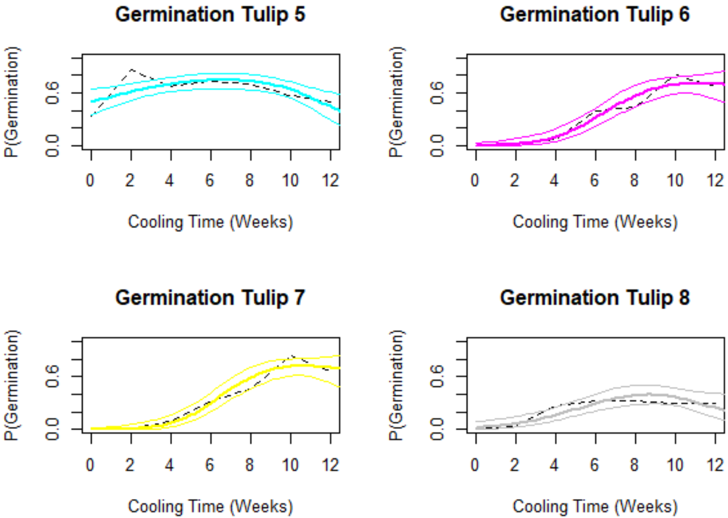
Spline1	Est	2.5 %	97.5 %
Pop1	4.523	2.659	6.489
Pop2	6.007	4.240	7.941
Pop3	9.377	7.087	12.041
Pop4	8.722	6.354	11.517
Pop5	1.454	-0.053	2.980
Pop6	11.626	7.570	16.975
Pop7	13.641	8.831	20.026
Pop8	5.977	3.281	9.316
Pop9	68.397	30.799	116.710
Pop10	11.185	8.047	15.101
Pop11	14.505	9.736	20.588
Pop12	0.000	-2.2e6	2.2e6

Spline2	Est	2.5 %	97.5 %
Pop1	-0.044	-1.477	1.602
Pop2	0.360	-0.799	1.644
Pop3	-0.468	-1.763	0.942
Pop4	0.692	-0.353	1.814
Pop5	-1.367	-2.357	-0.402
Pop6	2.184	1.149	3.291
Pop7	2.345	1.290	3.478
Pop8	-0.076	-1.202	0.984
Pop9	12.487	6.811	20.505
Pop10	2.119	0.938	3.487
Pop11	1.078	0.074	2.104
Pop12	0.000	-8.9e5	9.9e6

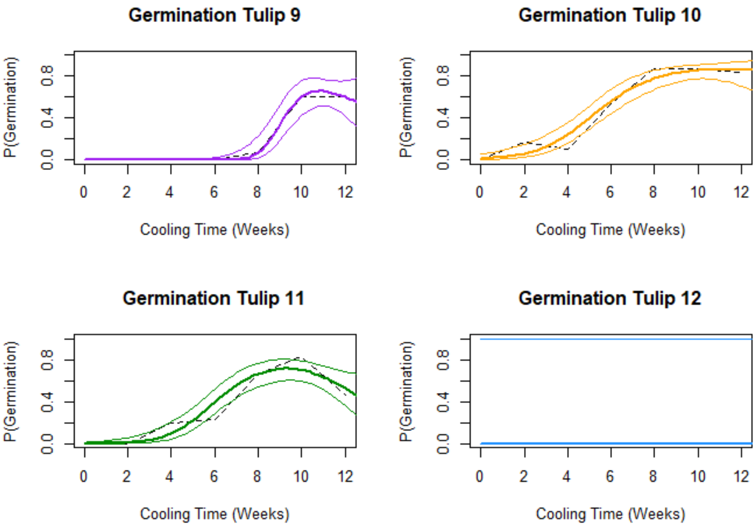
Easier to interpret with Pictures



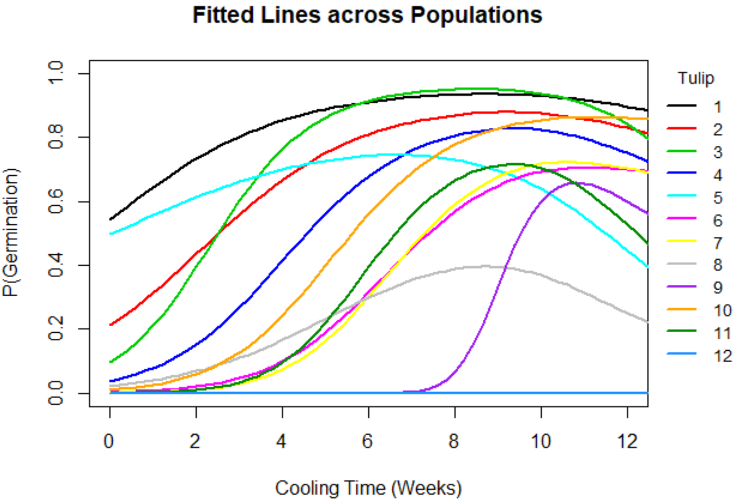
Cont'd


Germination Tulip 5
Germination Tulip 6
Germination Tulip 7
Germination Tulip 8
P(Germination)
0.0
0.6
0 2 4 6 8 10 12
Cooling Time (Weeks)

Cont'd

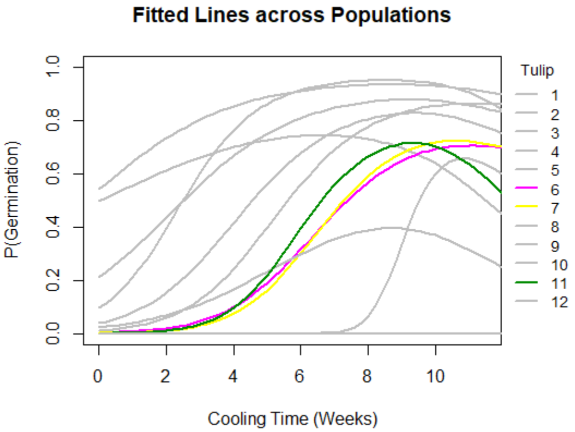

Germination Tulip 9
Germination Tulip 10
Germination Tulip 11
Germination Tulip 12
P(Germination)
0.0 0.4 0.8
0 2 4 6 8 10 12
Cooling Time (Weeks)

Are the probabilities per cooling time the same across populations?



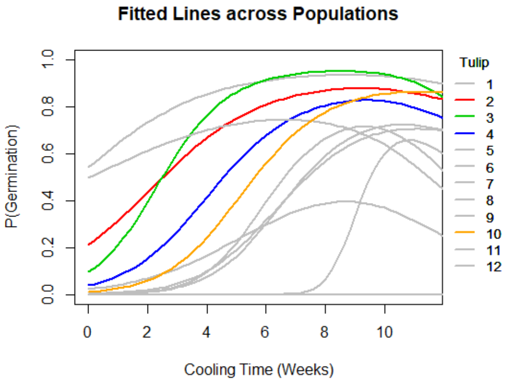
Which ones are the same?

- Tulips 6, 7, and 11 all take a while to grow and have a small optimal window. (Likelihood Ratio test χ^2 p-val: 0.104)
- These are at high risk under climate change conditions



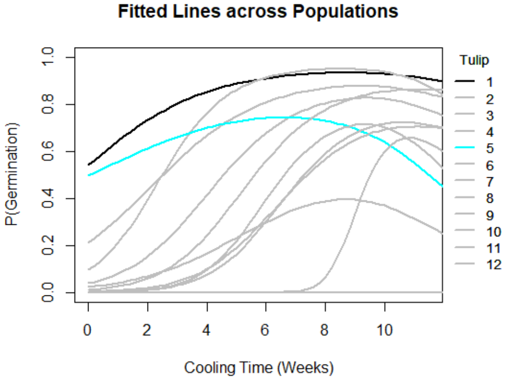
Which ones are the same? (2)

- Tulips 2,3,4 and 10 all need moderate cooling time but stabilize.
- Tulips 2 & 3 (LRT P-val: .125) and 4 & 10 (LRT p-val .104) were not statistically different.
- These are at lower-moderate risk conditions



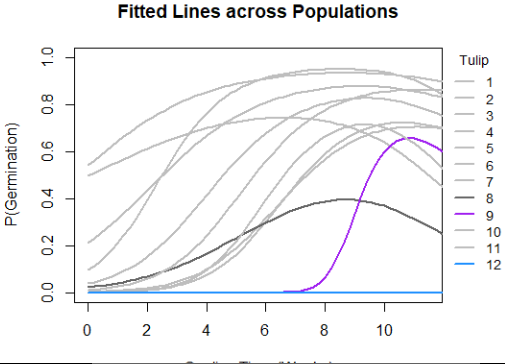
Which ones are the same/different? (3)

- Tulips 1 and 5 don't need much cooling time.
- However, tulip 5 needs peaks well before 10 weeks. This one may improve with climate change
- These have very low risk



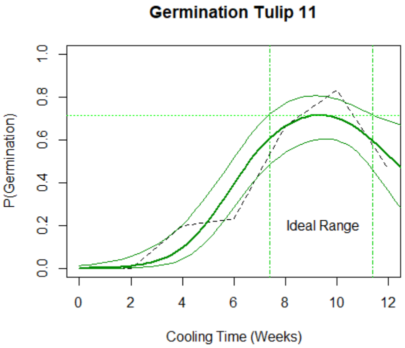
Which ones are the same/different? (3)

- Tulips 8 and 9 take longer periods to grow, but are somewhat stable.
- Tulip 8 is at lower risk with climate change, but is difficult to germinate
- Tulip 9 may not grow at all with climate change conditions!
- Tulip 12 may not be affected by cooling conditions.



"Ideal" Chilling time

- "Ideal" Chilling time: where predicted probability is the highest
- Ideal Range will be where upper 95% confidence interval contains predicted maximum.

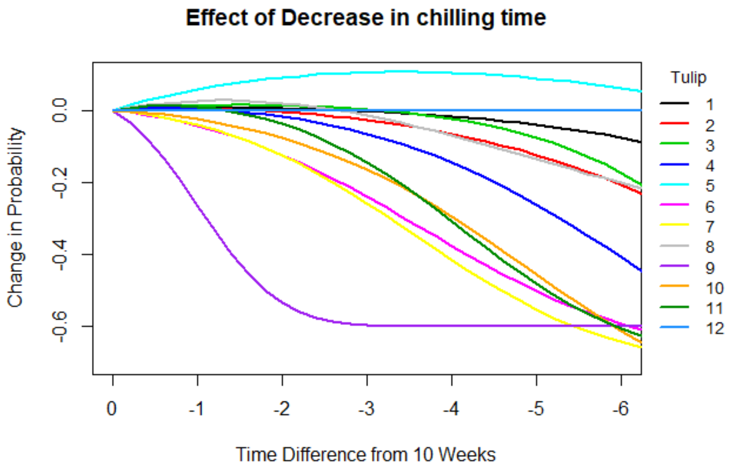


Q2 Results

Population	Maximum (Weeks)	Ideal Range
1	8.8	5+ Weeks
2	9.2	6+ Weeks
3	8.4	6-11 weeks
4	9.4	7+ Weeks
5	6.6	3-9 Weeks
6	11.0	8+ Weeks
7	10.6	8+ Weeks
8	8.8	6+ Weeks
9	10.8	10+ Weeks
10	11.0	8+ Weeks
11	9.4	7-11 Weeks
12	NA	Unknown

Effect of decrease chilling time

Typical chilling time is 10 weeks. What if this time was lowered?



Tulips 6,7,10,11 will take the largest hit, 2,3,4,8 are fine until chilling time goes below 6 weeks, 5 actually increases

Difference at 8 weeks

Pop	Est Diff	2.5%	97.5%
1	0.058	-0.412	0.528
2	-0.048	-0.424	0.329
3	0.233	-0.187	0.653
4	-0.120	-0.458	0.218
5	0.423	0.120	0.727
6	-0.537	-0.868	-0.207
7	-0.566	-0.901	-0.231
8	0.082	-0.258	0.422
9	-3.054	-4.701	-1.407
10	-0.523	-0.918	-0.127
11	-0.178	-0.490	0.134
12	0.000	-575.986	575.986

- Note: These differences are on the log-odds scale. We can only interpret the sign.
- If cooling period is shortened by 2 weeks by climate change, 4 are statistically worse.
- Tulip #5 actually germinates better!

Difference at 6 weeks

Pop	Est Diff	2.5%	97.5%
1	-0.244	-0.893	0.406
2	-0.498	-1.017	0.022
3	-0.347	-0.904	0.210
4	-0.793	-1.258	-0.328
5	0.486	0.070	0.901
6	-1.583	-2.132	-1.035
7	-1.767	-2.374	-1.160
8	-0.314	-0.801	0.173
9	-9.165	-14.535	-3.795
10	-1.531	-2.090	-0.972
11	-1.289	-1.856	-0.721
12	0.000	-795.831	795.831

- This scenario would be problematic: 6 Tulips are now statistically worse.

Risk under Climate Change

Tulip	Stablizes?	Peak Growth	Climate Risk
1	Yes	5+ Weeks	Very Low
2	Yes	6+ Weeks	Low
3	Somewhat	6-11 Weeks	Low
4	Somewhat	7+ Weeks	Moderate
5	No	3-9 Weeks	Very Low
6	Somewhat	8+ Weeks	High
7	Somewhat	8+ Weeks	High
8	Yes	6+ Weeks	Low
9	Somewhat	10+ Weeks	Very High
10	Yes	8+ Weeks	High
11	No	7-11 Weeks	Moderate
12	Unknown	Unknown	NA

Conclusion

Goals of the study

- In summary, we concluded that Tulips 6, 7, 9, and 10 are at most risk during climate change conditions.
- We used logistic regression to:
  1. Quantify Uncertainty
  2. Compute Predicted Probabilities
  3. Answer inference-based questions
- Most tulips peaked around 6-10 weeks
- Climate change clearly limits growth of tulips

Shortcomings

- Some spline fits did okay job at matching the data
- Extrapolation is still dangerous. Knot placement matters
- Accuracy of 80% is decent, but could be better.

Next Steps

- More data around 8-12 week range to improve fit
- Further investigate Tulip #12. Poor batch?
- Consider using different types of splines/algorithms

References

https://www.healthytravelblog.com
12/9/2013 *5-healthy-days-in-amsterdam*