Survival of the Fittest: Variable Selection on Agricultural Data from the Galápagos Islands

Michael Bostwick

April 19th, 2018

Table of Contents

Background

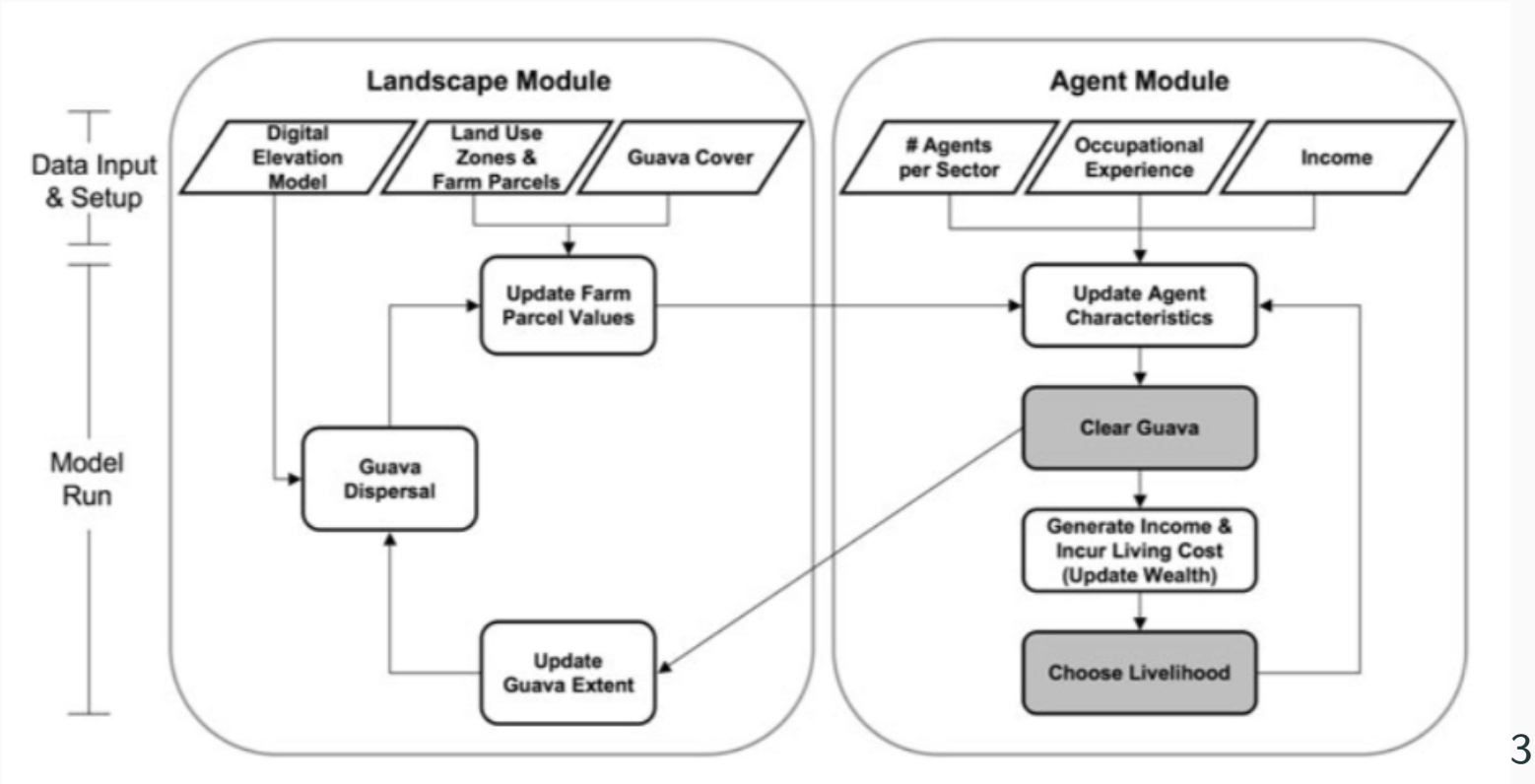
Galápagos Islands

[1]


[2]

[1] https://canconf.com/map-of-south-america-including-galapagos-islands/
[2] https://www.flickr.com/photos/pancholp/albums

Agent-Based Simulation



[3]

[3] Miller, B. W., Breckheimer, I., McCleary, A. L., Guzmán-Ramirez, L., Caplow, S. C., Jones-Smith, J. C., & Walsh, S. J. (2010). Using stylized agent-based models for population environment research: a case study from the Galápagos Islands. Population and environment, 31(6), 401-426.

Census of Farms

Consejo de Gobierno del Régimen Especial de Galápagos

Ministerio de Agricultura, Ganadería, Acuacultura y Pesca

CENSO DE LAS UNIDADES DE PRODUCCIÓN AGROPECUARIA (UPA) DE GALÁPAGOS

Número de cuestionario utilizados en la UPA ____ de ____

Cantón _____ ____

Parroquia _____ ____

Encuestador _____ ____

Ley de Estadística
Obligatoriedad y confidencialidad de la Información

Art. 20.- Todas las personas naturales o jurídicas domiciliadas residentes, o que tengan alguna actividad en el país, sin exclusión alguna, están obligadas a suministrar, cuando sean legalmente requeridas, los datos o informaciones exclusivamente de carácter estadístico censal, referentes a sus personas y a las que de ellas dependan, a sus propiedades, a las operaciones de sus establecimientos o empresas, al ejercicio de su profesión u oficio, y, en general, a toda clase de hechos y actividades que puedan ser objeto de investigación estadística o censal.

Art. 21.- Los datos individuales que se obtengan para efecto de estadística y censos son de carácter reservado; en consecuencia, no podrán darse a conocer informaciones individuales de ninguna especie, ni podrán ser utilizados para otros fines como de tributación o conscripción, investigaciones judiciales y, en general, para cualquier otro uso distinto del propiamente estadístico o censal. Sólo se darán a conocer los resúmenes globales, las lotizaciones y, en general, los datos impersonales.

CAPÍTULO 1. CARACTERÍSTICAS GENERALES

Características de la Unidad de Producción Agropecuaria (UPA)

1. Nombre de la Unidad de Producción Agropecuaria: _____
   Vía/Calle Km/Número

2. Ubicación de la Unidad de Producción Agropecuaria: _____

3. ¿Cuenta la Unidad de Producción Agropecuaria con línea telefónica? ____
   1. Sí - Registre el número → _____
   2. No - Continúe

4. Condición Jurídica de la Unidad de Producción Agropecuaria: ____
   1. Individual - (Registre el código y pase a la pregunta 6)
   2. Sociedad de hecho sin contrato legal
   3. Sociedad legal (Corporación, Sociedad Anónima, etc)
   4. Institución Pública
   5. Otra

5. Si es Sociedad o Institución Pública. ¿Cuál es el nombre de la empresa/institución? _____

6. ¿Dispone la UPA de página web? ____
   1. Sí - Registre el URL → _____
   2. No - Continúe

7. ¿Dispone la UPA o la persona productora de correo electrónico? ____
   1. Sí - Registre el e-mail → _____
   2. No - Continúe

8. ¿Dispone la UPA de alguna(s) de la(s) siguientes autorizaciónes? ☐ 1. Agencia de Regulación y Control Sanitario del MSP ☐ 2. Agencia de Bioseguridad de Galápagos (ABG) ☐ 3. Otra ¿Cuál? _____ ☐ 4. Ninguna

Características de la Persona Productora (Persona natural, principal responsable de las operaciones en la UPA)

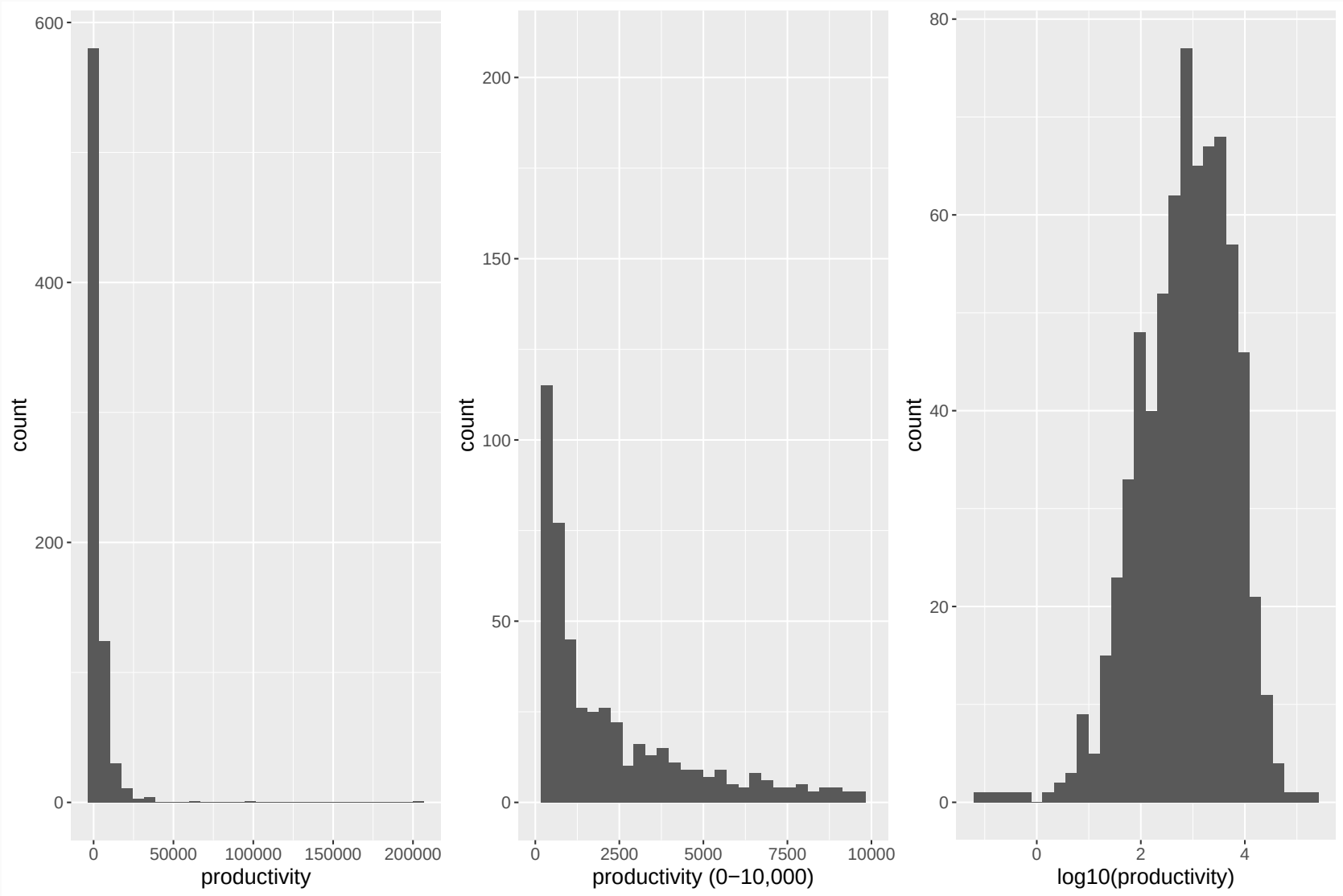
Nombres Primer Apellido Segundo Apellido

Data

Data description

Model 5 different response variables with 200+ possible predictors

Data transformation

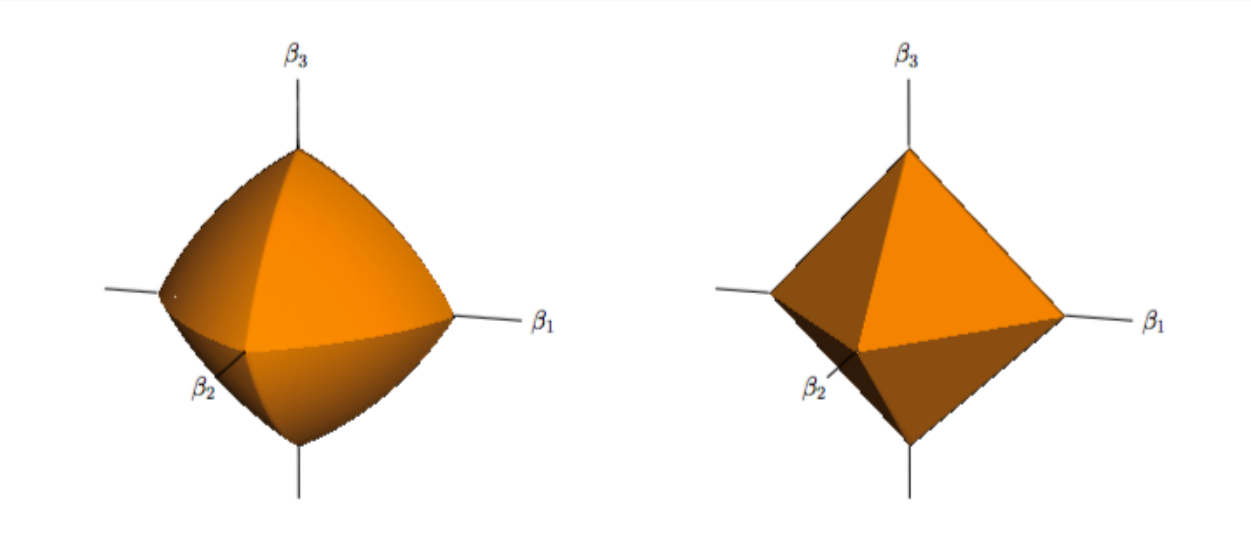


Modeling

Elastic Net

Formulation

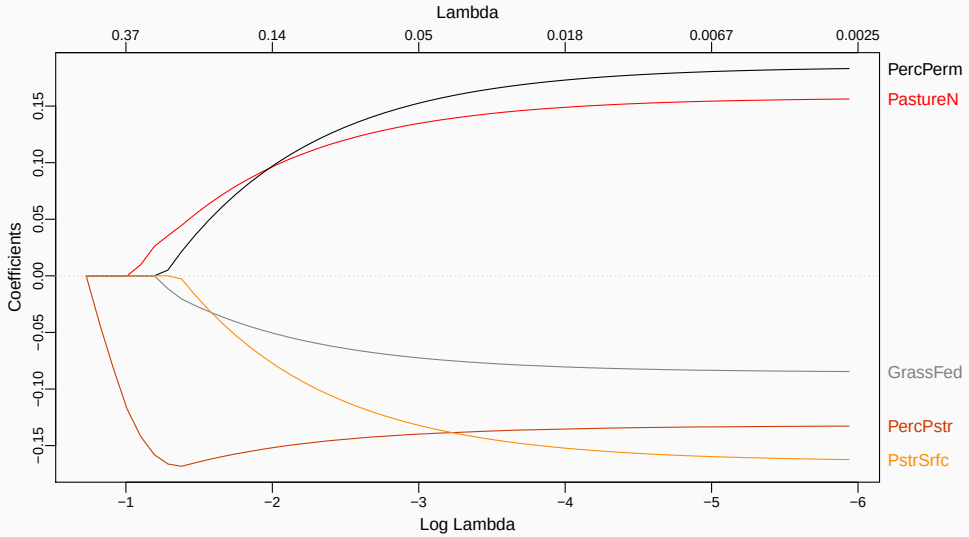
$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda[\underbrace{(1-\alpha)||\beta||_2^2}_{\text{Ridge}} + \underbrace{\alpha||\beta||}_{\text{Lasso}}]$$



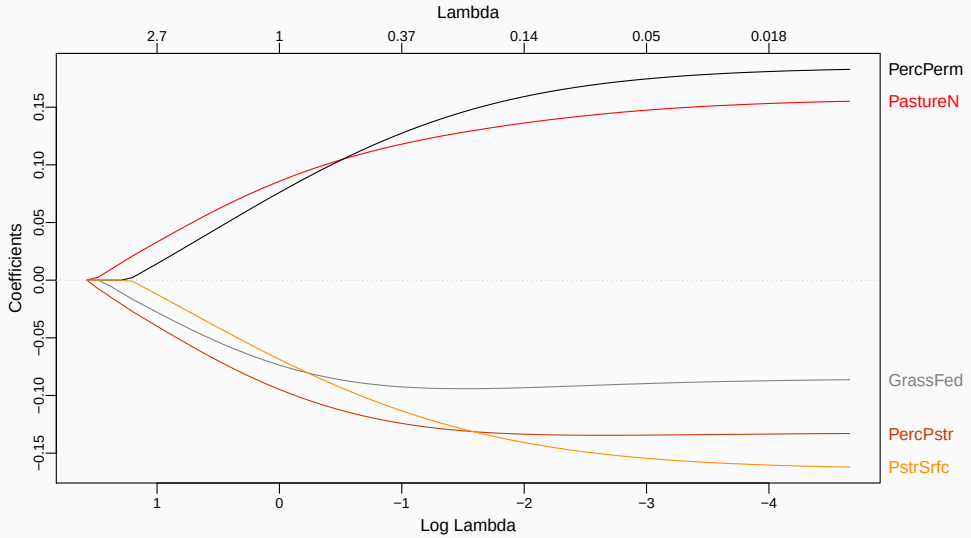
[4]

[4] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations. CRC press.

Elastic Net Coefficients



Alpha = 1



Alpha = 0.1

Forward Stepwise

1. Start with null model
2. Fit p simple linear regression models and pick one with lowest residual sum of squares (RSS)
3. Search through remaining $p - 1$ variables and add one that best improves RSS
4. Repeat Step 3 until all variables have been added to the model
5. Choose optimal model using information criterion, Bayesian Information Criterion (BIC) in this case

$$BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2)$$

Results

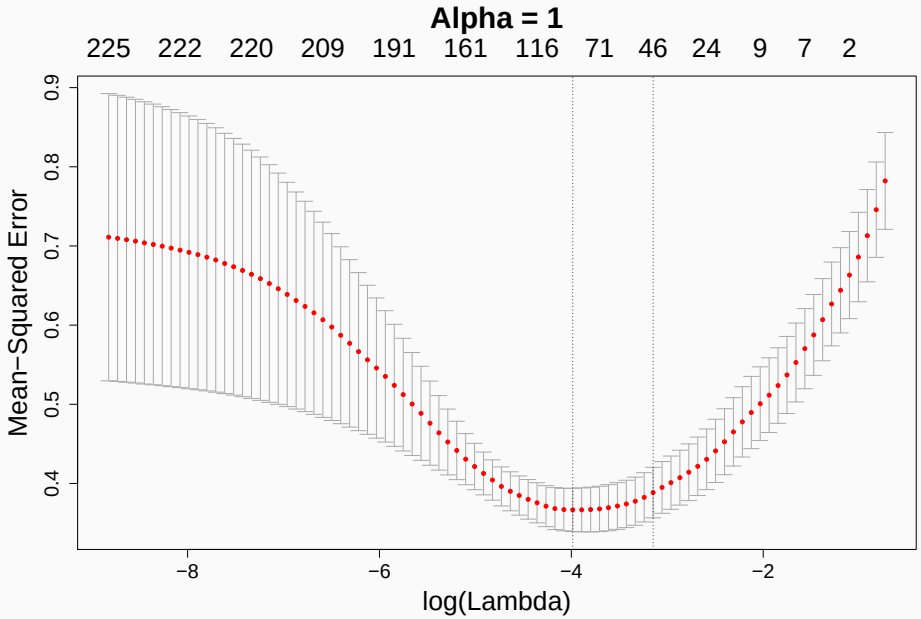
Response Variables

- Productivity
- Net Income
- Number of Workers Supported
- Invasive Species
- Land Use Choices

Productivity Model - Elastic Net

First 5 variables to enter model

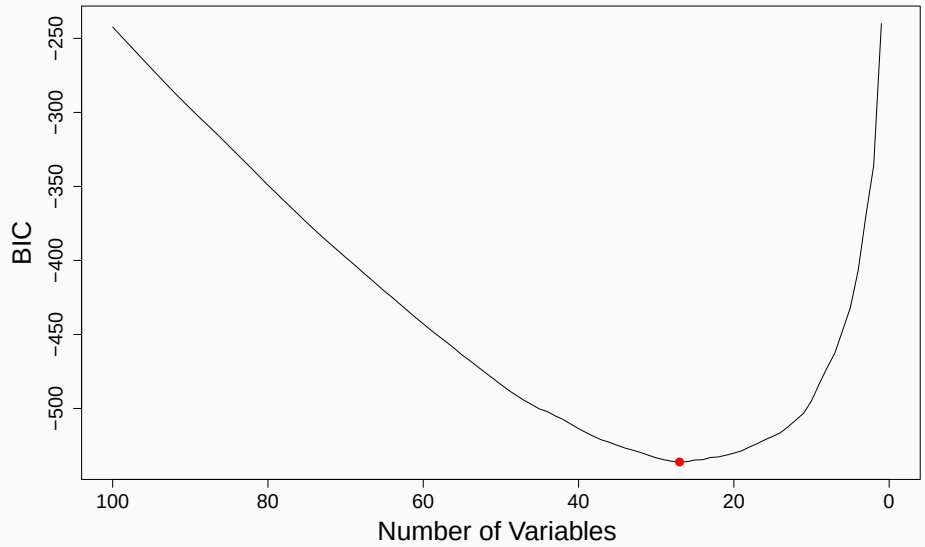
1. Percent pasture land (-)
2. No pasture land (+)
3. Percent of feed from grass (-)
4. Percent permanent crops (+)
5. Surface area of pastures (-)



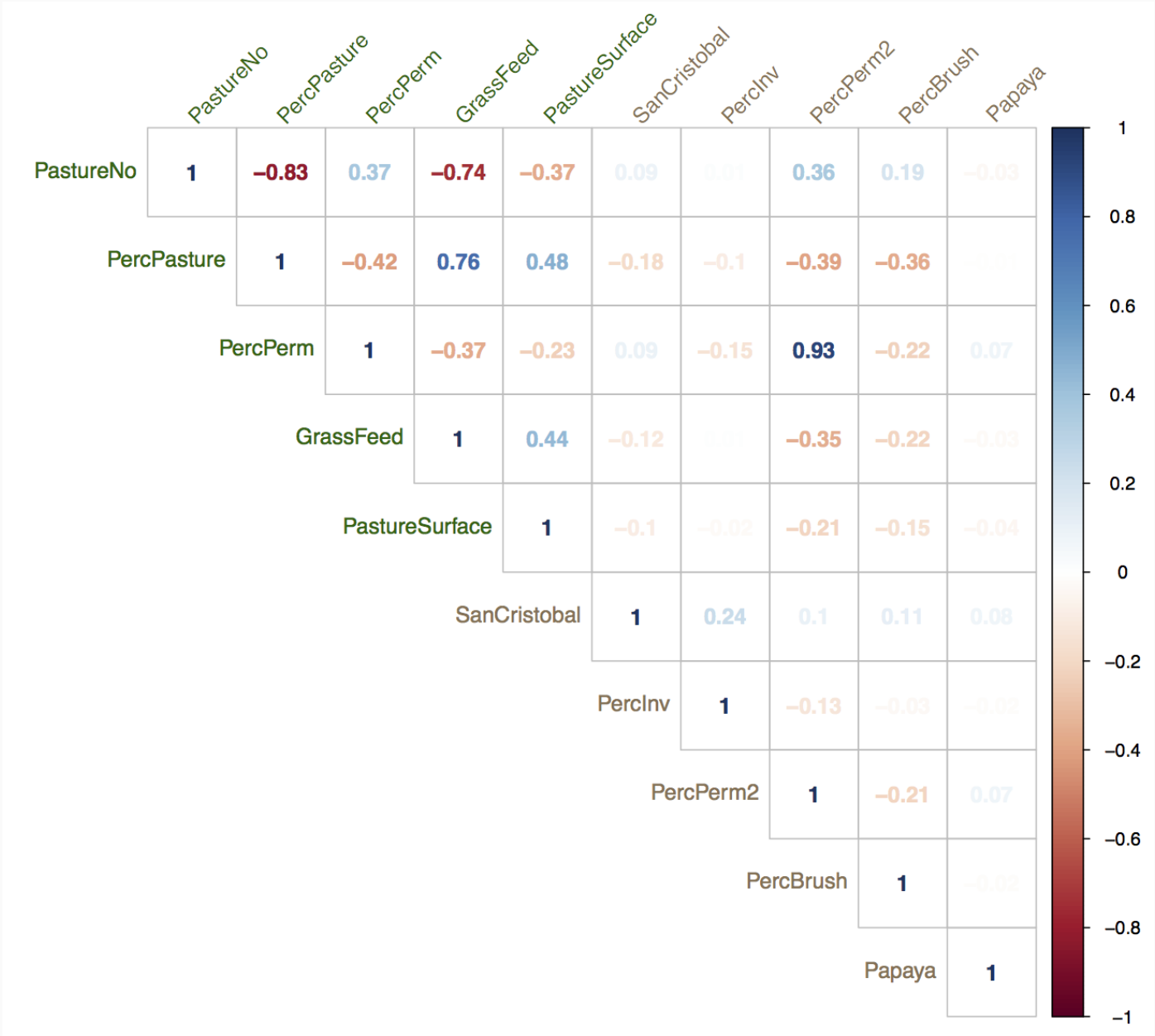
Productivity Model - Forward Stepwise

First 5 variables to enter model

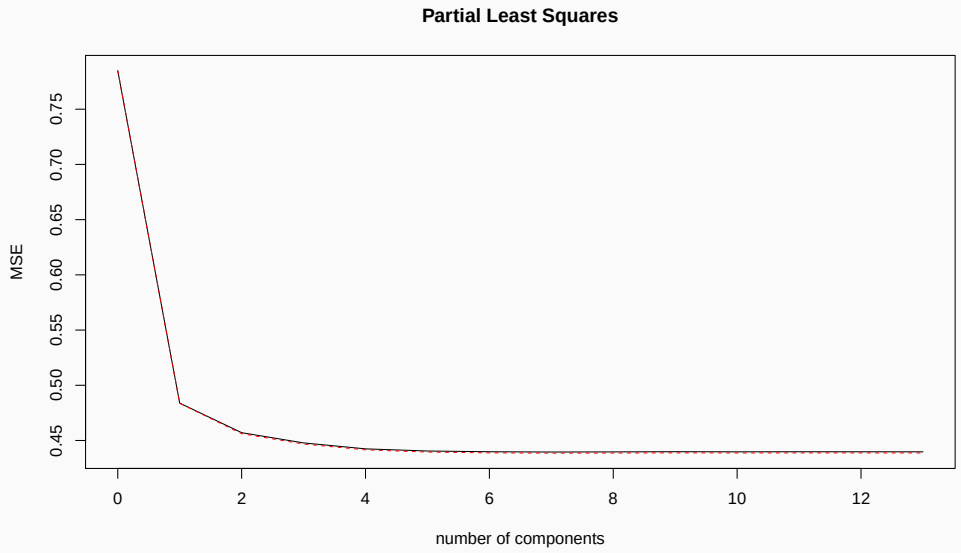
1. San Cristobal Island (+)
2. Percent invasive species (-)
3. 2 year growth in permanent crops (+)
4. Percent brush (-)
5. Growing papaya (+)

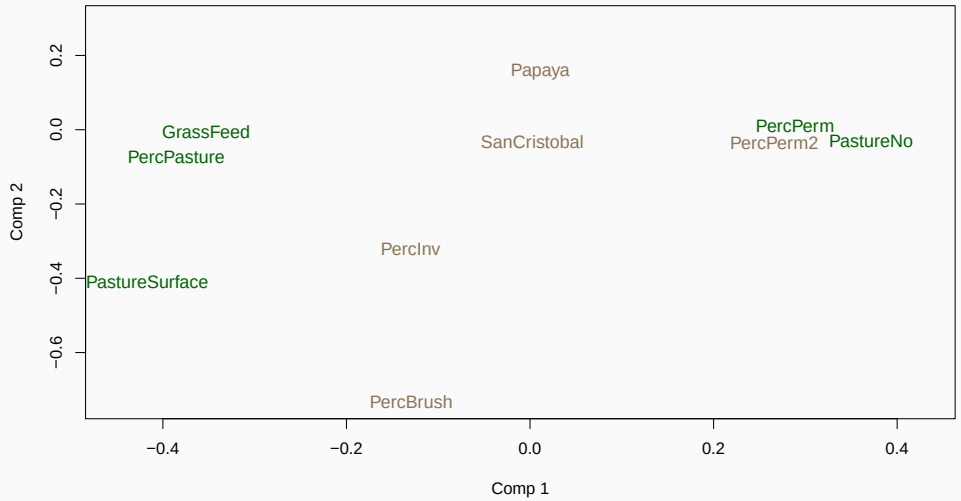


Productivity Model - Comparison



Productivity Model - Comparison





Questions?