A Longitudinal Analysis of the Pattern of Change Over Time for Soybean Yield (1990 – 2013) By Country

Emma Pham

Lauren M. Matheny, PhD, MPH

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

INTRODUCTION

- Soybeans are one of the most important crops and have a variety of uses as a high protein crop, oil crop[1] and animal feed.
- Crop yield is highly influenced by environmental factors, such as atmospheric carbon dioxide level (CO_2) and fluctuations in temperature [2].
- **The purpose of this analysis is to investigate if there is a difference in the pattern of soybean yield (hg/ha) per country over time using annual average temperature and annual pesticide use as predictors.**

METHODS

- A sample of 797 records across 34 countries with 24 repeated measurements (1990 – 2013) of soybean yield (hg/ha) from the Food and Agriculture Organization and World Data Bank were used for this longitudinal analysis.
  - The dependent variable, soybean yield (hg/ha), is treated as a continuous variable.
  - Countries with multiple (>1) average annual temperatures or countries with <20 occasions were excluded.

Assessing Pattern of Change Over Time (1990 – 2013)

- A spaghetti plot for visualizing the trajectories of soybean yield for each of the 34 countries was utilized to aid in the identification of patterns in soybean yield from 1990 to 2013 and to assess overall variation in soybean yield for all countries.
- After an unstructured saturated means model was generated for descriptive purposes to assess covariance structure and the pattern of soybean yield over time, the Intraclass Correlation (ICC), or the proportion of total variance attributed to between-person differences, was generated by an empty means, random intercept only model.
- Fixed and random effects of time were added iteratively to a longitudinal mixed model using a bottom-up approach and assessed at each iteration for best model fit. Time was centered at the first year (1990).

Assessing Model Fit

- Likelihood Ratio Test (-2 Log-Likelihood), Akaike's information criterion (AIC), Bayesian Information Criteria (BIC), or pseudo-R^2 (proportion reduction in each variance component as a result of adding fixed effects) were used to assess model fit at each addition of a fixed or random effect of time.

Normality

- Normality of soybean yield residuals were assessed using several visuals and Shapiro-Wilk's test.

RESULTS

- **Figure 1:** Soybean yield for many of the countries seem to fluctuate sharply from year to year, as shown by the high peaks and sharp drops over time.
  - Several countries differed by having lower yield (between ~5,000 hg/ha and 10,000 hg/ha) and appear the most stable/consistent in terms of yield over time.
- The ICC indicates that the proportion of variance due to Level 2 between-person differences is 77.38% and the proportion of variance due to Level 1 within-variation is 22.62%, which provides justification to proceed with a longitudinal model.
- **Table 1:** The Random Linear Model (addition of random linear time slope)'s -2LL, AIC, BIC values outperformed the Fixed Linear Time Model. A smaller -2LL/AIC/BIC indicates better fit.

Figure 1. *Soybean Yield (n=797) by Country from 1990 – 2013.*

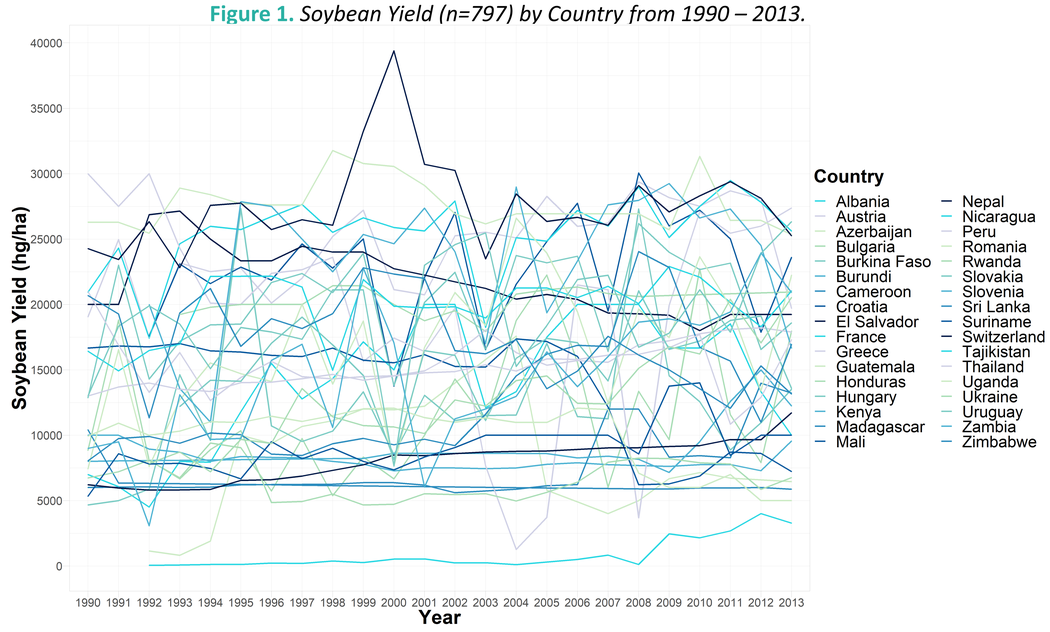


Table 1. *Fixed and Random Effects with Associated Model Fit Statistics for Model Comparison.*

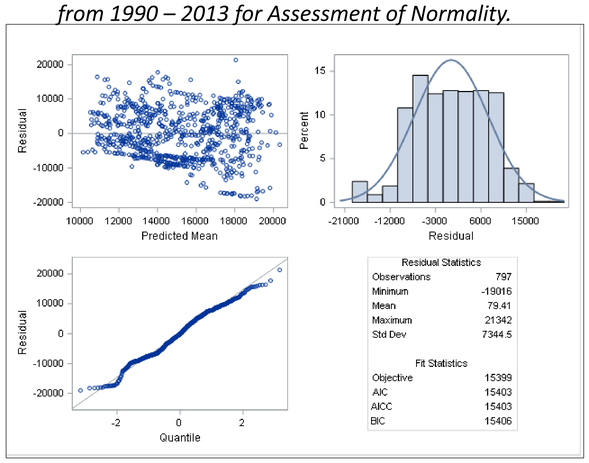
Parameter Estimates		*Fixed Linear Time, Random Intercept Model*			*Random Linear Time Model*			*Fixed Quadratic, Random Linear Time Model*		
		Estimate	*SE*	*p <*	*Estimate*	*SE*	*p <*	*Estimate*	*SE*	*p <*
Model for the Means										
γ_{00}	Intercept	13,833	1,177	<.0001	13,791	1,255	<.0001	13,304	1,280	<.0001
γ_{10}	Linear Time Slope	119	18	<.0001	122	40	0.0042	247	76	0.0012
γ_{20}	Quadratic Time Slope	-------	-------	-------	-------	-------	-------	-5	3	**0.0527**
	Pseudo-R^2									**0.0040**
Model for the Variance										
τ^2_{U0}	Random Intercept Variance	45,007,540	11,212,935	<.0001	51,726,856	13,176,366	<.0001	51,795,336	13,191,472	<.0001
τ^2_{U1}	Random Linear Time Slope Variance	-------	-------	-------	43,471	13,001	0.0004	43,822	13,082	0.0004
	Covariance Between Random Interecept and Random Linear Effect of Time	-------	-------	-------	-539,129	314,269	0.0863	-544,618	315,572	0.0844
σ^2	Residual Variance	12,478,348	639,282	<.0001	10,381,192	543,582	<.0001	10,339,388	541,770	<.0001
REML Model Fit										
	Number of Parameters	4			6					
	-2LL	15,410.0			**15,326.2**			-------		
	AIC	15,414.0			**15,334.2**			-------		
	BIC	15,417.1			**15,340.3**			-------		

Table 2. *Parameter Estimates of Fixed Effects for Random Linear Time Model with Time Invariant Predictors.*

Parameter Estimates	*Random Linear Time Model*		
	Estimate	*SE*	*p <*
Effect			
Intercept	22,370	2,919	<.0001
Linear Time Slope	133	40	**0.0023**
Average Temperature (°C)	-492	145	**0.0012**
Pesticide Use (tons of active ing.)	0.0152	0.0200	0.4487

Bolded p-values indicate significance (p<0.05).

Figure 2. Residuals for *Soybean Yield (n=797) by Country from 1990 – 2013 for Assessment of Normality.*



RESULTS (contd.)

- **Table 1:** The non-significant p-value (0.0527) for the quadratic fixed effect and very small pseudo-R^2 (effect size) (0.0040) contributes to the decision to exclude the quadratic fixed effect in the mixed model.
- The final model chosen is the:

Random Linear Longitudinal Conditional Model

Level 1: Soybean $Yield_{ti} = \beta_{0i} + \beta_{1i}(year_{ti}) + e_{ti}$

Level 2: Intercept: $\beta_{0i} = \gamma_{00} + \gamma_{01}(\text{Average Annual Temperature})_i + \gamma_{02}(\text{Annual Pesticide Use})_i + U_{0i}$

Linear Time: $\beta_{1i} = \gamma_{10} + U_{1i}$

- **Figure 2:** It appears that the residuals in the scatterplot follow a random pattern, which is indicative of normality and the histogram appears to be approximately normal due to the bell-shape. The residuals in the QQ plot seem to follow the normal line, although there are deviations from normality at the tail ends.
  - A Shapiro-Wilk's test indicates normality is not met (p <0.0001). This may be due to deviations in the residual tail ends seen in the QQ plot causing the non-normality. Based on the body of evidence, it appears that normality is reasonably met.

DISCUSSION

- Increases and fluctuation in soybean yield over time (Figure 1) serve as justification for the inclusion of fixed and random effects of time in the final model to capture the pattern of soybean yield.
- The variances and covariances from the marginal V matrix from the Random Linear model appear to decrease as years pass. There is an inherent autocorrelation in that years closer together are more correlated than years farther apart.
- Soybean yield increased on average by 133 hg/ha per year (Table 2) from 1990 to 2013. A variance for the random linear year slope was added (along with its covariance with the random intercept), resulting in a significant improvement in model fit, $-2\Delta LL(\sim X) = 15309.5$, $p < 0.05$, **indicating significant individual differences in the linear rate of increase in soybean yield across time.**
- There was no relationship (p>0.05) between annual pesticide use and soybean yield. There was a relationship between average annual temperature and soybean yield (p=0.0012); for every degree Celsius increase, soybean yield decreases by 492 hg/ha, while adjusting for time and temperature (Table 2).

SAS CODE

```
PROC MIXED DATA=work.soybeans1 COVTEST NOCLPRINT NAMELEN=100 IC
METHOD=REML;
    CLASS country year;
    MODEL yield = time avg_temp pesticide_use / SOLUTION DDFM=Satterthwaite;
    RANDOM INTERCEPT time / G GCORR V VCORR TYPE=UN SUBJECT=country;
    REPEATED year / R TYPE=VC SUBJECT=country;  ODS OUTPUT
CovParms=CovRandLin InfoCrit=FitRandLin; *Save for pseudo-R2 and LRT;
ESTIMATE "Intercept at year=1990 Time=0"    intercept 1 time 0;
*continue effect coding for remaining 22 occasions; RUN;
```

REFERENCES

1. https://www.sciencedirect.com/science/article/pii/B9780128013090000124
2. https://doi.org/10.1029/2021jg006304