Taylor’s power law and the stability of crop yields

Thomas F. Döring a,⁎, Samuel Knapp b, Joel E. Cohen c

a Department of Agronomy and Crop Science, Albrecht-Thaer-Institute of Agricultural and Horticultural Sciences, Faculty of Life Sciences, Humboldt University Berlin, Albrecht-Thaer-Weg 5, 14195 Berlin, Germany
b Department of Plant Sciences, Technical University of Munich, Emil-Ramann-Straße 2, 85350, Freising-Weihenstephan, Germany
c Laboratory of Populations, Rockefeller University & Columbia University, 1230 York Avenue, Box 20, New York, NY 10065, USA

A R T I C L E   I N F O

Article history:
Received 19 May 2015
Received in revised form 5 August 2015
Accepted 6 August 2015

Keywords:
Coefficient of variation
Crop
Finlay–Wilkinson regression
Stability
Taylor’s power law

A B S T R A C T

Taylor’s power law (TPL) describes the empirical relationship \( \sigma^2 = \alpha \mu^b \) where \( \sigma^2 \) are sample variances and \( \mu \) are sample means in subsets of data in a data set. Equivalently, TPL states that the logarithm of the sample variance is a linear function of the logarithm of the sample mean across different subsets of data. Here we show that crop yields follow this relationship in several published data sets from varied situations. We show that TPL is frequently, but not always, valid for various factors structuring the data including varieties, crop species, trial environments or countries. We propose that the residuals from the linear regression of \( \log(\sigma^2) \) against \( \log(\mu) \) can be used as a measure of stability, called POLAR (Power Law Residuals). We compare POLAR stability with other commonly used measures of stability, and show that POLAR stability offers an advantage over some frequently used stability measures.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In 1961, the British ecologist and entomologist Roy Taylor reported a series of observations that showed surprisingly similar patterns in data from several insect groups, earthworms and other invertebrates as well as plant viruses and fish (Taylor, 1961). For all these groups, population counts or other measures of density showed a power-law relationship between the sample variance \( \sigma^2 \) and the sample mean \( \mu \). This relationship, \( \sigma^2 = \alpha \mu^b \), became known as Taylor’s Power Law (TPL), or Taylor’s Law of fluctuation scaling (Cohen, 2013), though the relationship had been published and used earlier by other researchers. Logarithmic transformation of the equation \( \sigma^2 = \alpha \mu^b \) results in a linear relationship: \( \log(\sigma^2) = \log(\alpha) + b \log(\mu) \). Taylor, who originally found this relationship for data from aerial sampling of aphids, suggested that aggregation is key to understanding the mechanisms underlying this power law and he interpreted \( b \) as an index of aggregation.

Subsequently, TPL was verified in hundreds of further species (e.g., Eisler et al., 2008), with data stemming from a wide variety of ecosystems, from bacterial cultures in a laboratory (Ramsay et al., 2012) to forests (Cohen et al., 2012; Cohen et al., 2013). TPL was also found to be valid for a tremendous range of non-biological phenomena including traffic of Internet routers (Duch & Arenas, 2006), transactions of the New York Stock Exchange, rainfall, and the printing activity of a large university printer (Eisler et al., 2008). Because of its ubiquity, TPL has attracted wide attention from empirical researchers and theorists of multiple, often unrelated disciplines.

As data demonstrating TPL are not restricted to ecological research, the original notion suggested by Taylor and co-workers that TPL follows from explicit behavioural mechanisms such as aggregation and migration has become untenable. An alternative explanation of TPL (Anderson et al., 1982) is that TPL has stochastic roots and that behavioural mechanisms are not necessary to generate data sets exhibiting TPL. In this view, TPL can result from demographic and environmental stochasticity. TPL can indeed be generated by very different population growth models such as the Lewontin-Cohen random walk (Cohen et al., 2013) and the exponential model (Cohen, 2013). While debate on stochastic vs. behavioural explanations of TPL continues (Kendal and Jorgensen, 2011; Arruda-Neto et al., 2012), it has recently been suggested that both perspectives are valid in that “the behavioral models have the potential to explain how the interactions of individuals could generate the population-level phenomenology summarized by the stochastic models” (Cohen, 2013).

In a related area of research, H. Fairfield Smith showed linear relationships between the logarithm of plot size in so-called ‘blank’ field experiments (without any treatments) and the logarithm of the variance of wheat yield per plot (Smith, 1938). Smith, however, was not interested in the relationship between means and variance...
per se, but his main interest was to use this relationship to find an optimal plot size (also see Taylor et al., 1999).

As far as we are aware, TPL has not yet been applied to or tested for crop yields. In this paper we demonstrate the validity of TPL for crop yields in various contexts and on various scales. Further, we introduce a new measure of yield stability which is based on TPL. We call this new index POLAR stability (based on Power LAW Residuals) and compare this approach with the behaviour of two commonly used stability indices, the coefficient of variation (CV) (Francis and Kannenberg, 1978) and the Finlay–Wilkinson (FW) regression slope (Finlay and Wilkinson, 1963).

Material and methods

Data sets and data filtering

Data sets were analysed to represent crop yield variation (1) at various levels including plot level (within field trials) (Seufert et al., 2012); (2) between environments (years and trial sites) (Jones et al., 2010; Vlachostergios et al., 2011); and (3) at a global level (national yield statistics from the Food and Agriculture Organization database, FAOStat [FAOSTAT, 2014].

The dataset from Seufert et al. (2012) was collected for a meta-analysis comparing the yields of conventionally and organically managed crops. The dataset contains observations from published studies where for each observation the mean yields and corresponding standard deviations were collected. The original dataset, containing 316 paired observations of conventional and organic yields, was downloaded from the supplementary files published online with the original study.

The data were then filtered in the following way. First, all data not convertible into units of ha$^{-1}$ were discarded (e.g., units of boxes or bales per ha, or units of kg per plant). Data in bushels per acre were converted as 1 bu/ac$^{-1}$ = 0.062777 t ha$^{-1}$ for maize and 0.06725 t ha$^{-1}$ for wheat and soybean. Second, the original dataset contained some double or multiple entries where means and variances were equal among different observations within studies. For each system separately, only one of the equal observations within studies was kept. The final dataset contained 268 observations from the organic system and 228 for the conventional system, i.e., a total of 496 data points from 33 crop species. Observations from maize, wheat, tomato and soybean dominated the dataset with 128, 80, 45 and 43 observations, respectively.

The second dataset is from a study on the adaptability of different lentil genotypes to organic farming, conducted in Greece (Vlachostergios et al., 2011). Twenty genotypes were grown in five environments (three years in one location and two years in another location). In each environment (i.e., year $\times$ location combination), the lentils were grown under both organic and conventional cropping management, in separate trials, and with three replicates each. Calculations are based on the reported means over three replicates.

In the third dataset, 19 genotypes of wheat, from two quality groups ( milling and feed), were grown at two organic sites (Sheepdrove and Wakelyns) and two conventional sites (Metfield and Morley) in the UK over three years (Jones et al., 2010); the published data are the means of three replicates per trial environment. For calculation of TPL, means and variances for each genotype are calculated across the environments. For the lentil and the wheat study, means and variances were calculated across (a) all organic environments ($n_{L} = 5$ for lentils, $n_{W} = 6$ for wheat); (b) all conventional environments ($n_{L,c} = 5$, $n_{W,c} = 6$); and (c) all environments ($n_{L,a} = 10$, $n_{W,a} = 12$).

The FAO dataset was downloaded in November 2014 from http://faostat3.fao.org/download/ For wheat, rice and potatoes, data were extracted for area (recorded in ha) and yield (recorded in kg ha$^{-1}$). These data were then filtered. First, only those countries with a complete set of the last ten years (2004–2013) were retained. Second, countries where average area between 2004 and 2013 was below 10,000 ha were excluded. In addition, to exclude data that indicated low reliability, we excluded countries in which yields were rounded to 1000 kg ha$^{-1}$ in two or more of 10 years and countries where yields were equal in consecutive years in two or more of 10 years. For wheat, rice and potatoes, these filters resulted in final datasets of crop yields from 90, 73 and 89 countries, representing 99.3%, 97.9% and 96.8% of the global area grown with these crops in 2013, respectively.

Calculations and statistical analysis

All analyses were performed with the programme R, version 3.0.0. Means ($\mu$) and variances ($\sigma^{2}$) were calculated per variety over all environments for the lentils and wheat dataset, and per country over all years for the FAO dataset. Subsequently, a linear regression was calculated for $\log_{10}$ of the variance over the $\log_{10}$ of the mean. In order to test the linear relationship between $\log(\mu)$ and $\log(\sigma^{2})$ we applied a forward selection procedure by starting with a linear model and subsequently adding higher order polynomials (quadratic and cubic terms) and evaluating the model fits by the Akaike information criterion (AIC), i.e., terms were added until the AIC increased (Burnham and Anderson, 2002). Cubic terms were never significant in any of the analyses.

Residuals from the linear regression models were tested for normality, skewness ($m_{3}$) and kurtosis ($m_{4}$). Deviation from normality was tested with the Shapiro–Wilks test (Royston, 1982b). Deviation of skewness from zero and deviation of excess kurtosis ($m_{4} − 3$) from zero were tested with $t$-tests on $n − 2$ degrees of freedom (Crawley, 2013); ‘excess kurtosis’ $= 0$ is equivalent to kurtosis $= 3$, since the normal distribution has a kurtosis of 3. In case of non-normal residuals, the variable $y = \log(\sigma^{2})$ was transformed using the Box–Cox transformation ($z = (y^{\lambda} − 1)/\lambda$) with maximum likelihood optimisation of $\lambda$ (Crawley, 2013), in order to quantitatively describe the behaviour of the residuals. Homoscedasticity was checked visually using the plot(model) function in R and with the Goldfeld–Quandt test (Goldfeld & Quandt, 1965) using the generic function in the R library lmtest. In addition, the relationship between $\log(\mu)$ and $\log(\sigma^{2})$ was tested with Spearman’s rank correlation coefficient as a non-parametric test not dependent on the assumptions of linear regression.

2. Results

Taylor’s power law: testing validity for crop yields at various scales

Here we show that TPL is sometimes but not universally valid for crop yields across various scales (across environments, and at a global level) and for various factors structuring the data (genotypes, crop species, countries, years).

For the dataset from Seufert et al. (2012), based on variances of crop yields mainly within field trials, there was a highly significant ($P<0.001$) positive correlation between $\log(\mu)$ and $\log(\sigma^{2})$ (Fig. 1a), whether the organic and conventional systems were analysed separately or together. Paired observations from the two systems (organic and conventional) of both means and variances are stochastically dependent. Strictly speaking, this would affect the validity of significance tests for the regression and also tests of assumptions. However, similar results were obtained when fitting regressions within systems separately (Fig. 1a), where means and variances were stochastically independent. According to the Goldfeld–Quandt test, the data were not significantly heteroscedas-
The relationship between $\log(\mu)$ and $\log(\sigma^2)$ for individual crop species with ten or more observations per system in the filtered dataset was much less robust than when all crops were analysed together (Table 1). For some crop species, namely maize, soybean, barley and spinach, the regression slope was not significantly different from zero. However, when individual observations were averaged per crop within the systems, resulting in 66 observations in total, the regression across all crop species of $\log(\sigma^2)$ against $\log(\mu)$ was highly significant (Table 1, Fig. 1b). For this analysis across crop species, there was no significant deviation from linearity regardless of whether the organic and conventional systems were analysed separately or together. Generally, high significance and low S.E. values for the regression slopes were mainly observed when the range of values spanned by $\log(\mu)$ was large (Table 1).

For the lentil dataset (Vlachostergios et al., 2011), there was a highly significant linear relationship between $\log(\mu)$ and $\log(\sigma^2)$ (Fig. 2, $P < 0.001$, df = 18) and no significant deviation from linearity ($P = 0.692$, df = 17). When the means and variances were calculated within management systems, the relationship between $\log(\mu)$ and $\log(\sigma^2)$ was evident for the conventional, but not the organic systems (Table 1). For neither of the two systems considered separately was there any significant quadratic term. However, significant heteroscedasticity ($P < 0.05$) was found for the conventional lentil data subset.

For the wheat dataset from Jones et al. (2010), variances and means were calculated across environments for each variety. There was a highly significant linear relationship between $\log(\mu)$ and $\log(\sigma^2)$ (Fig. 2). The wheat dataset showed a significant linear relationship between $\log(\mu)$ and $\log(\sigma^2)$ in the organic, but not in the conventional, systems (Table 1). The Goldfeld–Quandt test did not detect significant heteroscedasticity in any subset of the data from Jones et al. (2010).

The non-significant TPL slope for the conventionally grown wheat yields from Jones et al. (2010) was robust: omitting any one of the six conventional environments from the dataset always resulted in non-significant TPL slopes; the slopes had a negative sign in five out of six cases and in the remaining case (when Metfield data from the second year were omitted), the slope was close to zero ($b = 0.09 \pm 1.14$). In comparison, the organic wheat yields from Jones et al. (2010) retained positive TPL slopes with omission of individual environments (analyses not shown). In addition, the TPL slopes of the conventional wheat remained non-significant when both groups of wheat cultivars (milling cultivars and feed cultivars, see original study) were analysed separately, whereas
the organic wheat data retained significantly positive TPL slopes for both milling cultivars \(b = 1.33 \pm 0.56\), \(df = 9\), \(P = 0.04\) and feed cultivars \(b = 2.53 \pm 0.92\), \(df = 5\), \(P = 0.04\).

Deviation from linearity was significant for the organic systems (significant quadratic terms, \(P < 0.05\)) but not the conventional systems, nor when both systems were considered together. When variances and means were calculated across varieties within each of the 12 environments, there was again a significant linear relationship.

For all three crop species from the FAOSTAT yield dataset, there were significant linear relationships between \(\log(\mu)\) and \(\log(\sigma^2)\) (Fig. 3, Table 1). Deviations from linearity of the regression lines of \(\log(\sigma^2)\) against \(\log(\mu)\) were not significant in any of the three crop species. For the wheat data, but not for the rice and potato data, the Goldfeld–Quandt test showed significant heteroscedasticity \(P < 0.05\).

**Consequences of Taylor’s Power Law for some commonly used yield stability measures**

In the FW regression approach to yield stability analyses (Finlay and Wilkinson, 1963), sometimes also called 'ecoregression' (Meinak et al., 1997), yields of individual genotypes are displayed in a scatter diagram against the environmental means, which are the means over a set of genotypes within each environment. These environmental means are used as a surrogate environmental index for the yielding ability or quality of the respective environment. If a genotype \((G)\) shows no interaction between genotype performance and environment, the slope \(\beta_G\) of the regression line of genotype \(G\) yields against the environmental means is 1. Genotypes that show relatively better performance in poor environments (defined as environments with lower mean yields) have slopes \(\beta_G < 1\), whereas genotypes with disproportionately higher yields in good environments (defined as environments with higher mean yields) have slopes \(\beta_G > 1\).

What are the consequences if, as our previous results indicate, variances scale with environmental means in accordance with TPL? For example gain yield data of genotypes of wheat from Jones et al. (2010) resulted in \(\log(\sigma^2) = -1.637 + 1.939 \times \log(\mu)\). Consequently, the confidence interval for yields of individual genotypes or individual varieties becomes broader as the environmental mean increases (Fig. 4). These increasing confidence interval could arise

---

**Table 1**

Leastsquares regression parameters for \(x = \log(\theta_0)\) (temporal mean) and \(y = \log(\sigma^2)\) for yield data \((\mu = \text{mean}; \sigma^2 = \text{variance}; \theta = \text{unit of original data}; \theta_0 = \text{the lower bound of data})\) from various sources (Org: organic; Con: conventional; the range Q is the difference between largest and smallest value of \(\log(-\mu)\); i.e., \(Q = \log(\max_{\theta}(\mu)) - \min_{\theta}(\log(\mu))\); \(\cdot \cdot \cdot : \text{ns} \neq 0.00 \cdot \cdot \cdot \); not significant; NA: not available; **Significance level: Linear term.**

<table>
<thead>
<tr>
<th>Crop</th>
<th>Source dataset</th>
<th>Crops calculated</th>
<th>df</th>
<th>100(0)</th>
<th>200(5)</th>
<th>300(10)</th>
<th>400(15)</th>
<th>500(20)</th>
<th>600(25)</th>
<th>700(30)</th>
<th>800(35)</th>
<th>900(40)</th>
<th>1000(45)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wheat</td>
<td>Senfert et al. (2012)</td>
<td>164</td>
<td>78</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Tomato</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maize</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potato</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spinach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apple</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other averages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wheat</td>
<td>FAOSTAT (2014)</td>
<td>88</td>
<td>88</td>
<td>0.68</td>
<td>0.66</td>
<td>0.63</td>
<td>0.58</td>
<td>0.53</td>
<td>0.49</td>
<td>0.45</td>
<td>0.40</td>
<td>0.35</td>
<td>0.30</td>
</tr>
<tr>
<td>Wheat</td>
<td>FAOSTAT (2014)</td>
<td>87</td>
<td>87</td>
<td>0.68</td>
<td>0.66</td>
<td>0.63</td>
<td>0.58</td>
<td>0.53</td>
<td>0.49</td>
<td>0.45</td>
<td>0.40</td>
<td>0.35</td>
<td>0.30</td>
</tr>
<tr>
<td>Wheat</td>
<td>FAOSTAT (2014)</td>
<td>86</td>
<td>86</td>
<td>0.68</td>
<td>0.66</td>
<td>0.63</td>
<td>0.58</td>
<td>0.53</td>
<td>0.49</td>
<td>0.45</td>
<td>0.40</td>
<td>0.35</td>
<td>0.30</td>
</tr>
</tbody>
</table>

**Fig. 3.** Relationship between \(x = \log(\theta_0)\) (temporal mean) and \(y = \log(\sigma^2)\) for national yield data of wheat (grey circles, dashed line), rice (black diamonds, continuous line) and potatoes (white squares, bold line). Data are from FAOSTAT for 2004–2013. Each point represents one country; mean and variances are over years. For statistics of regression lines see Table 1. Low residual potato yield variability (i.e. high temporal yield stability) is indicated for two examples, Viet Nam and South Africa (black arrows, negative residuals from the regression line, i.e., POLAR < 0). Angola and Spain show positive residuals (POLAR > 0), i.e. low yield stability (grey arrows).
Fig. 4. Schematic representation of FW regression, based on grain yield data of wheat from Jones et al. (2010). Yields of individual genotypes are plotted against the environmental means, exemplified by the genotype with the highest and lowest
mean yield, Deben (black circles) and Thatcher (white diamonds) respectively. Values on the x-axis are environmental means, i.e. yields are averaged across all
genotypes in each environment. Around the bold line ($y=x$), the two fine lines ($y=x+1$ and $y=x-1$) show hypothetical genotypes for which there is no interaction
between genotype (G) and environment (E). Traditional stability analyses interpret
slopes $\beta_C$ of genotype-specific lines deviating from $\beta_C = 1$ as an indication of G x E
interactions. However, if variances scale with the environmental means, as pre-
picted by Taylor’s power law (TPL) and as demonstrated for this wheat dataset,
the confidence interval for values of individual genotypes, indicated by the dashed
lines, becomes broader with an increasing environmental mean. Both axes are on
the original scale of measurement, not on the logarithmic scale as in TPL.

if genotypes with below-average yields have FW regression slopes
of $\beta_C < 1$ while genotypes with above-average yields have slopes
$\beta_C > 1$. Alternatively, increasing confidence intervals could arise if
$\beta_C = 1$ for all genotypes, and the absolute value of residuals $d_C$ from
the FW regression lines (i.e., the difference between the yield of an
individual genotype and the yield predicted by the regression
across environmental means) increased with increasing environ-
mental mean. The variance of the residuals $d_C$ is often used as a
further stability measure along with the FW regression slope $\beta_C$
(Eberhart and Russell, 1966).

The wheat data support the hypothesis of increasing FW regres-
sion slope $\beta_C$ with increasing individual mean yield for genotype
$G$. Values of the genotype-specific slopes $\beta_C$ calculated for the 19
varieties were highly positively correlated with mean yields of the
varieties (Adj. $R^2 = 0.716$, df = 17, $P < 0.001$), while the variances ($\sigma^2_c$)
of the residuals were negatively, but less strongly, correlated with the mean (Adj. $R^2 = 0.250$, df = 17, $P < 0.05$).

A further commonly used measure of yield stability is the coef-
cient of variation, $CV = \sigma / \mu$, where $\sigma$ is the sample standard
deviation and $\mu$ the sample mean. This stability index is popu-
lar with ecologists (Tilman et al., 1998; Tilman, 1999; Anderson
et al., 2013) and agronomic researchers and is frequently calcu-
lated along with other stability measures (Francis and Kannenberg, 1978; Smith et al., 2007; Dehghani et al., 2008). However, if TPL
holds, as it does for (some) crop yield data, then the CV may change
non-linearly with increasing mean because $CV = \mu b^{2-1} g^{1/2}$, where $a$
and $b$ are the regression parameters (intercept and slope) of the TPL
log–log regression and $g$ is the basis of the logarithm. Consequently,
if $b > 2$ (as for conventional and organic wheat yields combined, in
the data of Jones et al., 2010; Table 1), CV increases with increasing $\mu$; if $b < 2$ (as in all remaining cases in Table 1), CV decreases with increasing $\mu$; and if $b = 2$, CV is constant and independent of $\mu$. The

FAO wheat data analysed in Table 1 have $b = 0.781 < 2$, hence CV
decrees with increasing mean yield $\mu$. The non-linear decrease of CV with increasing mean is also evident for the independent FAO
dataset of the wheat yields in 1994–2003 ($n = 90$ countries) (Fig. 5).

Characterisation and interpretation of residuals from TPL
regression

For the data from Seufert et al. (2012), residuals from the linear
TPL regression were non-normal, with significantly negative skewness
and significantly positive excess kurtosis (Table 2). Negative skewness of residuals was also found for the lentil dataset
(Vlachostergios et al., 2011), the wheat dataset (Jones et al., 2010),
and for wheat and potatoes of the FAO dataset. Deviation from
normality of the residuals was significant in half of the cases, but
excess kurtosis was mostly not significantly different from zero.
For the data from Seufert et al. (2012), where individual means and
variances were based on different numbers $n$ of measurements,
residuals from the linear TPL were not significantly correlated with
$n$ (Adj. $R^2 < 0$, $P > 0.5$, df = 378). Non-normality of residuals is a viola-
tion of the assumptions of linear regression. However, Spearman’s
rank correlation coefficient as a non-parametric way to measure the
strength of the relationship between log($\mu$) and log($\sigma^2$) showed
significant positive relationships between the two variables in all
cases (Table 2) in which linear regression had been found to be
significant. Further, with Box–Cox transformation of the data, nor-
mality of residuals could be obtained in most but not all cases.

The residuals from the linear TPL regression can be interpreted
as a measure of yield stability (Fig. 3); positive residuals indicate
that the log($\sigma^2$) is larger than expected for a given log($\mu$), and neg-
ative residuals indicate that the log($\sigma^2$) is smaller than expected,
i.e. that the yield is less variable. We call these residuals POLAR
(Power Law Residuals). POLAR stability is measured in units of
log(variance), not variance per se, because POLAR are not resi-
duals on the variance scale from the power law, but residuals from
the linear regression of log(variance) as a function of log(mean).

There was no significant correlation between these POLAR sta-
bility values and the original yield means $\mu$ for any of the data
analysed (Adj. $R^2 < 0.01$ and $P > 0.6$ for all three crop species in the
FAO dataset, as well as for the datasets from Jones et al. (2010) and
Vlachostergios et al. (2011); for data from Seufert et al. (2012), Adj.
$R^2 = 0.003$; $P = 0.129$).

In comparison to the results obtained with non-transformed
data, Box–Cox data transformation for the normalisation of resi-
duals changed the stability ranking of individual entries, though
these changes were small. In the organic subset from Vlachostergios et al. (2011), 12 out of 20 lentil genotypes retained their POLAR stability ranking, and 6 changed by only one rank; in the joint dataset (organic and conventional), 4 genotypes did not change their ranking, whereas 10 changed by one rank. Similarly, for the conventional wheat data from Jones et al. (2010), 14 wheat genotypes retained their POLAR stability ranking. In the potato data from FAO, 64 of 89 countries changed POLAR stability ranking by three or fewer ranks following data transformation.

3. Discussion

Validity of Taylor’s power law for crop yields

For several diverse datasets of crop yields, \( \log(\sigma^2) \) increased with increasing \( \log(\mu) \); in several cases, this relationship was linear. Given the wide applicability of TPL (Eisler et al., 2008), this finding is not surprising. However, it raises the question why TPL has not been shown for crop yields before (to our knowledge).

One plausible reason is that TPL tends to become more visible when the range of \( \log(\mu) \) is large. Taylor et al. (1988a,b) recommended that “the range of values of \( \log(\mu) \) in the regression form of the power law... should be as large as practicable.” (also see Taylor et al., 1988a,b.) Though not all datasets analysed here conform to this suggestion (e.g., wheat from organic environments in Jones et al., 2010), in general, the data with the largest overall range (from Seufert et al. (2012)) point into this direction (Table 1). If a larger range does render TPL more likely to become evident in crop data, then the chance of observing TPL will be low when comparing treatments or genotypes with similar means. Small ranges of means in crop yield data, and associated low detectability of relationships between means and variances may also be the reason why logarithmic transformation of yield data, though suggested early (Finlay & Wilkinson, 1963) and also used more recently (Piepho et al., 2014), is currently not common. In datasets with a relationship of means and variances, the assumption of homogeneity of variances may be achieved through a log-transformation. However, when crop yields are log-transformed TPLs are likely to disappear e.g., when the original wheat data from Jones et al. (2010) were log-transformed prior to the calculation of means and variances, the relationship of means and variances was non-significant (Spearman’s rank correlation: \( P = 0.90 \)).

A second reason why TPL has not been shown in crop yields so far may be that many measurements of yield are required to do so. For example, the unexpected negativity of the TPL slope for oats yield data reported in Table 1 disappeared when data were excluded from the analysis that were based on small numbers of yield measurements for each estimate of a mean and a variance. Taylor et al. (1988a,b) suggested that when testing TPL, at least 15 observations (here, measurements of yield) should be available for each estimate of a mean and a variance, and at least 5 pairs \( (\mu, \sigma^2) \) should be available for the regression of \( \log(\sigma^2) \) against \( \log(\mu) \). Taylor et al. (1988a,b) provided little statistical justification for these recommendations. Most estimates of means and variances reported in this paper are based on fewer than 15 measurements of yield.

Further, Taylor et al. (1998a,b) suggested that TPL is most reliable when sampling is conducted consistently, i.e. when the same statistical population is being sampled. The study by Fairfield Smith (1938) (and previous work cited therein) indicated that plot size is a major factor determining variance of yield. Consistent sampling, e.g., in terms of identical plot size in all environments, is likely to be the case with the data from Vlachostergios et al. (2011) and from Jones et al. (2010), but within the data from Seufert et al. (2012)
and within the FAO data, compilations are based on different data collection methods. Despite this, TPL fits well for both the Seufert et al. data and the FAO data.

In view of the limitations of sample size and sampling consistency, we regard the present results as (possibly) the first, exploratory work rather than the last, definitive word on crop yields and TPL. Further research will need to investigate how TPL in crop yield data depends on numbers and groupings of observations, as well as on methods of crop yield measurement.

A further important point highlighted by our analysis is the non-normality of residuals for TPL log–log regression of some of the crop yield datasets (Table 2). When the residuals are not normal, or when any other assumption made in linear regression is violated, we can still fit the model to the data by least squares, but the confidence intervals of intercept and slope and the significance level assigned to the linear correlation are no longer valid. (Ordinary least squares regression also assumes that the values of the x-coordinate, here \( \log(\mu) \), are known with no sampling or other error. This assumption is violated in all of our data, though the effect of this violation is likely to be small.) However, a non-parametric test that does not rest on the assumptions made for linear regression (Spearman’s rank correlation, Table 2) reveals significant relationships between \( \log(\mu) \) and \( \log(\sigma^2) \) in all cases where ordinary linear regression had shown significance. More caution is due when inferences are made based on estimates of variance (see below).

We regard TPL (and the associated analysis of skewness and kurtosis of residuals) mainly as a useful description of the relationship between mean and variance in crop yield data. The approach presented here provides opportunities of investigating effects of agricultural system design, e.g. comparisons of organic and conventional cropping management. Specifically, the analysis of the wheat data from Jones et al. (2010) suggests that under high-input (conventional) conditions, cultivars with high mean yields have lower than expected yield variability, since the associated TPL slope is not significantly positive. In contrast, the same varieties tested under organic conditions show a significantly positive TPL slope. Although these results need to be interpreted with caution because only two sites per management system were included in the field trials, the findings are plausible and in accordance with existing literature on the interactions between wheat genotypes and cropping system (Murphy et al., 2007). In particular, our results indicate that wheat breeding may have succeeded in combining high mean yields with low yield variability, but that these gains depend on the yield stabilizing effects of high inputs typically found in conventional crop management. However, why similar results were not found in the lentil dataset is open to speculation.

**Consequences of Taylor’s power law for yield stability measurement**

When TPL holds for crop yield data, as is often the case according to our results, some commonly used yield stability measures, such as the FW regression slope \( \beta_C \), the variance of the residuals, \( \sigma_C \), from the FW regression slope and the coefficient of variation (CV), need to be interpreted with caution.

The FW regression slopes \( \beta_C \) (or alternatively, the absolute residuals from the FW regression) are likely to increase with increasing means (Fig. 4) when the TPL slope is positive. In most cases, \( b > 0 \). In Table 1, \( b < 0 \) only for oats in the data of Seufert et al. (2012) and conventional wheat in the data of Jones et al. (2010). According to our analysis, an FW regression slope \( \beta_C \neq 1 \) does not necessarily result from \( G \times E \) interactions. If genotypes \( i, j \) have different mean yields \( (\mu_i \neq \mu_j) \), then TPL with slope \( b \neq 0 \) is incompatible with all genotypes showing the same yield stability as measured by \( \beta_C \) or absolute residuals from the FW regression.

Either the FW regression slopes of the two genotypes must differ to satisfy TPL, or the absolute residuals from FW regression must differ between the two genotypes.

This is also evident when considering some known biometric relationships. According to Becker (1981), the variances \( \sigma^2_C \) of individual genotypes \( G \) across environments (‘places’) \( p \) are a function of the FW regression slope \( \beta_C \) and the variance \( (\sigma^2_C)_p \) of the deviations \( d_C \) from the FW regression, namely, \( \sigma^2_C = \beta_C^2 \sigma^2_C + 2(\beta_C - 1)^2\sigma^2_p + (\sigma^2_C)_p = \alpha^2 p^2 \beta_C^2 + (\sigma^2_C)_p \). Here \( \alpha^2 \) is the variance of the effects of the places \( p \), and \( \sigma^2_p \) is taken to be the same for all genotypes. Thus, if \( \sigma^2_C \) increases with increasing \( \mu_C \) (as in TPL), then the means \( \mu_C \) must correlate positively with the FW regression slopes \( \beta_C \) or the deviations \( (\sigma^2_C)_p \) or both. As a consequence, differences in FW regression slopes or absolute residuals from the FW regression among genotypes may not always result from breeding, e.g. selection for stability, but may just be typical distributions of yields that result in variances scaling with the means according to TPL.

Further, Fig. 5 showed that the CV can decrease systematically with increasing mean yield in line with earlier work on wheat yield data (Taylor et al., 1999). We can specify that CV systematically decreases with increasing mean in crop yield data following TPL with \( b > 2 \), and that CV increases with increasing mean yield if \( b < 2 \). If CV is used as an index of yield stability, genotypes (or any treatments) with higher means will automatically have greater stability as assessed by smaller CV (if \( b < 2 \)) or smaller stability as assessed by large CV (if \( b > 2 \)). However, it is often desirable to separate yield stability from mean yields in order to detect biological synergies and trade-offs between mean performance and stability of performance.

As the ratio of standard deviation to mean, the CV is designed to make standard deviations more directly comparable across different means. However, when TPL holds with \( 0 < b < 2 \), the CV ‘overcompensates’ the scaling effect of the mean, so that higher means systematically lead simultaneously to higher variances and lower CVs. Accordingly, in a study comparing several indices of yield stability with a principal component analysis (Dehghani et al., 2008), the CV lay half-way between the mean and various other indices of yield stability.

Thus, both CV and FW regression stability are related to the mean of the treatments if TPL holds. Even if TPL is not strictly valid e.g. because of non-linearity, as long as variances are positively correlated with the means as shown by Spearman’s rank correlation (Table 2), both CV and FW regression stability are still related to treatment means. These facts affect attempts to test the hypothesis that there is a trade-off between yield and stability as the observed relation is probably more due to statistical relationships than the actual behaviour of the treatments under test. For example, in plant breeding and variety evaluation, a stability index strongly influenced by the mean would not allow selection of truly stable genotypes. This problem is particularly relevant when means differ widely (Fig. 5), e.g. when comparing the yields of different cropping systems or crop species. For these comparisons, a number of approaches to measure stability have been developed (Piepho, 1998). In practice, however, the CV has often been used (Rao and Willey, 1980; Mesterházy, 1995; Berzsenyi et al., 2000; Smith et al., 2007; Küchenmeister et al., 2012), especially in agronomic research, though mostly in conjunction with other stability parameters.

**POLAR as a new stability index**

Residuals from the TPL linear regression (POwer LAW Residuals or POLAR) can be interpreted as a measure of yield stability, with lower stability (relative to all samples with that mean yield) indicated by more positive POLAR values, and higher stability (relative
to all samples with that mean yield) indicated by more negative POLAR values.

In our data, there was no significant correlation between POLAR stability and mean yields, so POLAR was independent of the mean. Whether POLAR stability will be independent of the mean yields in general depends on the error structure of the relationship between log(σ²) and log(μ). If the residuals on the log(σ²) scale are homoscedastic, then POLAR stability and mean yields will be independent, fulfilling an important desideratum for stability measures.

Further, the POLAR stability index is easy to calculate e.g., with commonly available spreadsheet programmes, and easy to interpret mathematically and graphically (Fig. 3). If POLAR values are available from different groups within the dataset, significance of differences in POLAR stability between these groups can be tested. Such tests may need to be non-parametric in case of non-normal distributions of POLAR values, as characterised in Table 2. Conventional statistical tests that assume residuals are normally distributed may not be appropriate in all cases. Further, POLAR can be regarded as a special case of a more general approach of stability analysis. This approach first builds a model to describe the relationship between mean and variance and estimates variances for each mean; it then calculates the differences of observed and estimated variance (i.e. the residuals) as a measure of stability.

Over the past few decades, multiple methods have been developed to measure yield stability (Becker, 1981; Becker and Léon, 1988; Gauch, 1992; Eghbal and Power, 1995; Piepho, 1998). These measures have advantages and disadvantages. So does POLAR stability. Which genotype is most stable strongly depends on the stability index used (e.g., Dehghani et al., 2008). Research on yield stability has therefore often attempted to demonstrate relationships among different stability measures (Becker, 1981). We suggest that future studies of the consistency of stability measures in ranking genotypes, crop species or agronomic treatments should include POLAR stability among other stability measures, because it takes account of the dependence of the variance on the mean.

A characteristic of TPL is the robustness across datasets uniformly collected, in particular with respect to the slope b. Thus, the TPL response itself may differ among different crop varieties. In this case, changing the ensemble (i.e., subset) of varieties that is tested for stability across a number of environments may change the ensemble of associated TPL gradients. In the realistic case that two subsets A and B of varieties are tested, with some varieties being shared among the two subsets, the common gradient b₀ for the subset A may be different from the common gradient b₀ for subset B (b₀ ≠ b₀). This will quite likely change the POLAR response of individual varieties, drawing into question its general robustness, since each application may deliver a different result. One way to increase the robustness is to make sure that a sufficiently great number of observations is included. Similar problems are observed for other stability measures (e.g., FW regression slope β); however, more extensive datasets need to be analysed using TPL to elucidate how far POLAR is, or is not, robust against changes of ensemble and variations in the number of observations.

4. Conclusions

High stability of crop yield is an important goal in agricultural production (Becker & Léon, 1988; Piepho, 1998; Annicchiarico, 2009). To generate reliable income each year, farmers seek varieties or crops with low fluctuations in yields over time at the farm level. Since the early days of professional plant breeding, researchers and breeders have sought to develop crop cultivars with high yield stability over time and in multiple locations, so that the genotype can be successfully grown over large areas and in many different environments (Mühleisen et al., 2014; Chamekh et al., 2015; Ingversen et al., 2015). At geographic scales larger than an individual farm, temporal yield stability is critical for ensuring food security and reducing the risk of food shortages (McKersie, 2015). Research has developed several indices to characterise yield stability in various contexts and data structures.

Taylor’s Power Law suggests a new stability index that can deal with one form of the dependence of the variance on the mean across large ranges of mean values, under assumptions spilled out above. This index (POLAR stability) is easy to calculate and interpret. Further, as shown for the dataset from Seufert et al. (2012), it is applicable to historical data where only variances and means are known, but not the original values. TPL highlights the need to be cautious in interpreting stability measurements of crop yields when means vary over large ranges, because some measures of yield fluctuations may be biased when TPL holds. In cases when different entities (such as cropping systems or crop species) with large differences in mean yields are to be compared in terms of their yield stability, it is therefore essential to check for systematic relationships between means and variances; ideally, enough data points should be collected to determine TPL slopes.

We have shown that TPL is widespread but not universal in crop yield data. From the discovery of TPL until to today, the potential underlying mechanisms that lead to the TPL relationship between means and variances have been the focus of numerous studies. It is clear that as in other datasets, the original proposition of aggregation behaviour as the driving force generating TPL is not valid for crop yields. Recently it has been shown that relatively simple, non-biological mechanisms could explain some (but not all) occurrences of TPL in biology (Cohen and Xu, 2013; Xiao et al., 2015). It remains to be investigated in which cases deviations from TPL in crop data result from the size and nature of the data, and in which, alternatively, systematic factors explain these deviations from TPL. While it is too early to spell out implications of TPL for agronomy in detail, benefits can be expected for both crop science and the study of TPL across the sciences from the wide application of the TPL approach to crop yield data.

Acknowledgments

We thank Frank Ellmer and Charlotte Kling for helpful comments during the preparation of this paper. JEC acknowledges with thanks the support of U.S. National Science Foundation grant DMS-1225529 and the assistance of Priscilla K. Rogerson.

References
