9-28-2021

Incorporating Molecular Markers and Causal Structure among Traits Using a Smith-Hazel Index and Structural Equation Models

Juan Valente Hidalgo-Contreras

Josafhat Salinas-Ruiz

Kent M. Eskridge

Stephen P. Baenziger

agronomy

MDPI

Article

Incorporating Molecular Markers and Causal Structure among Traits Using a Smith-Hazel Index and Structural Equation Models

Juan Valente Hidalgo-Contreras [1,*], Josafhat Salinas-Ruiz [1], Kent M. Eskridge [2] and Stephen P. Baenziger [3]

[1] Campus Córdoba, College of Postgraduates in Agricultural Sciences, Km 348 Carretera Córdoba-Veracruz, Amatlán de los Reyes, Veracruz 94946, Mexico; salinas@colpos.mx
[2] Department of Statistics, University of Nebraska-Lincoln 340 Hardin Hall, Lincoln, NE 68410, USA; keskridge1@unl.edu
[3] Department of Agronomy and Horticulture, University of Nebraska-Lincoln 362D Plant Sciences Hall, Lincoln, NE 68410, USA; pbaenziger1@unl.edu
* Correspondence: jvhidalgo@colpos.mx; Tel.: +52-2711784176

Abstract: The goal in breeding programs is to choose candidates that produce offspring with the best phenotypes. In conventional selection, the best candidate is selected with high genotypic values (unobserved), in the assumption that this is related to the observed phenotypic values for several traits. Multi-trait selection indices are used to identify superior genotypes when a number of traits are to be considered simultaneously. Often, the causal relationship among the traits is well known. Structural equation models (SEM) have been used to describe the causal relationships among variables in many biological systems. We present a method for multi-trait genomic selection that incorporates causal relationships among traits by coupling SEM with a Smith–Hazel index that incorporates markers. The method was applied to field data from a Nebraska winter wheat breeding program. We found that the correlation and the relative efficiency increased for the proposed Smith–Hazel indices when the total causal information among traits was accounted for by the vector of weights (b), which includes the causal path coefficients in the causal matrix (Λ). On the other hand, when selection was based on a primary trait, for example yield, the proposed SI increased the mean yield of the best 28 (Top 10%) genotypes to 7%.

Keywords: selection index (SI); structural equation modeling (SEM); yield components; multi-trait; causal relationship

Citation: Hidalgo-Contreras, J.V.; Salinas-Ruiz, J.; Eskridge, K.M.; Baenziger, S.P. Incorporating Molecular Markers and Causal Structure among Traits Using a Smith-Hazel Index and Structural Equation Models. *Agronomy* **2021**, *11*, 1953. https://doi.org/10.3390/agronomy11101953

Academic Editor: Harbans Bariana

Received: 5 June 2021
Accepted: 17 September 2021
Published: 28 September 2021

1. Introduction

Novel food production technologies are needed to tackle the increased demand for food in a world in which population is expected to reach 9 billion people by 2050. The production gains achieved through conventional breeding methods are gradually declining. Producing more food requires the development and implementation of new genetic technologies. The goal of breeding programs is to choose candidates that produce offspring with the best phenotypes. In conventional selection the best candidate is selected with high genotypic values (unobserved) under the assumption that these genotypic values are related to the observed phenotypic values. Selection index (SI) is a tool that helps to select individuals by considering more than two quantitative traits simultaneously. The SI is a linear combination of different traits, in which each trait is weighted according to its importance [1]. Selection indices were first proposed by Smith [2] and Hazel [3] as a linear combination of the phenotypic traits of interest, in which the corresponding weight for each trait is obtained by maximizing the correlation between the phenotypic and the genotypic merits with respect to the weights. This index is called a Smith–Hazel index or an optimum index [4]. To estimate the vector of coefficients for the optimum index, one must know the economic trait for each trait and the phenotypic and genotypic variance-covariance

matrices among the traits. Unfortunately, there is no strict rule for setting the economic weights.

There are two general types of applications of SI. The first is single trait improvement, in which there is one trait of interest and other related traits are used to increase the efficiency of selection. The second type is to improve more than one trait simultaneously, which is also called multiple-trait improvement, where the main concern is to assign economic weights for each trait. In addition to the Smith–Hazel multiple trait index, other indices have been proposed, such as the base index, where economic weights are used as index coefficients: the restricted index and the non-weighted multiplicative index [4]. A common application of these indices has been applied to yield and yield components, among other quantitative traits.

The use of molecular markers (MM) in genetic studies has gained enormous attention in recent years, since markers may be used to explain a proportion of the total genetic variance. In this context, incorporating MM into selection indices to improve the estimation of breeding values has been proposed by several authors, either by adding a single-marker [5,6] or by adding k molecular markers [7]. A SI was developed to incorporate k molecular markers as an indirect trait, by regressing the breeding value over the molecular marker values (marker scores). In that SI, the essential steps are: (i) predict the genetic value, known as a marker score, by regressing phenotypes on marker information; and (ii) combining the marker score with phenotypic information and the SI to make the final predictions of the genetic merit [7].

In all the selection indices described above, it is possible to use several traits simultaneously, since they are based on the Smith–Hazel index theory. However, none of them use the causal relationships that exist between the traits. In agriculture, most of the traits of interest are direct or indirect functions of other traits. For example, in plant breeding, grain yield is a function of several intermediate traits. A unidirectional relationship between yield and yield components in small grains was established [8] (Figure 1). The path diagram describes the sequential development of yield components in small grains. For example, the spikes per square meter (SPSM) influence the number of kernels per spike (KPS), which in turn affects the kernel weight (KW), which is also affected by SPSM.

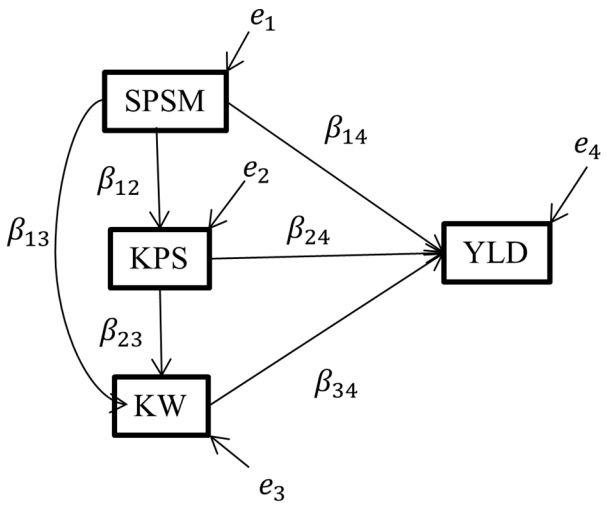

Figure 1. Path diagram showing a sequential relationship between yield (YLD) and yield components in small grains. SPSM = spikes per square meter, KPS = kernel per spike, and KW = kernel weight. Arrows represent the direction of the influence of variables. The β's represent the path coefficients. The e's represent the error term associated with each trait. Own elaboration based on [8].

Causal relationships among traits have been well established in biology for many years [9]. Specifically, the causal structure among yield components is well known. However, selection is still conducted ignoring this causal structure. Therefore, the purpose of this research was to make use of the Smith–Hazel index and to incorporate structural equation modeling (SEM) theory to improve the predictability of the model for selection response.

Multivariate statistical models (MSM) can handle several variables simultaneously. However, when it is known that there is a causal association among these traits, a MSM does not take advantage of this causal information. In this context, structural equation modeling (SEM) can consider the causal relationships among traits, to improve the predictability of the model.

Structural equation modeling (SEM) is a powerful statistical tool that is useful for understanding complex relationships among variables by considering the causal structure among the traits of interest. SEM models that account for biological causal relationships among traits can be useful in multi-trait selection strategies [10]. SEM is a modern version of path analysis, which was originally proposed by Wright [9] and has been widely applied in the social sciences, but not widely used in the plant sciences. Despite its lack of use in the plant sciences, SEM has potential applications in this field, such as for the analysis of yield components, genetics, and multi-environment studies. SEM was applied to characterize the genetic architecture in multivariate systems modeling the causal relationship among phenotypes [11]. Some applications of SEM in crop sciences have been to explore the causal relationship in grain yield components [12,13], as well as in other areas of plant sciences [14,15]. SEM was used to model genotype using environmental interaction [13,16]. In plant science, a few papers have been published using SEM in the analysis of yield and yield components [12–14,17,18].

SEM was used to adapt quantitative genetic models to model causal relationships between phenotypes, and also to show the statistical consequences when the association between two traits was analyzed in terms of standard multi-trait models (MTM), which ignored the causal association among traits [11]. The authors suggested that accounting for the causal information that is presented in many biological systems would be a more realistic way of addressing them [11].

SEM integrates path analysis, a system of simultaneous equations and factor analysis [19]. SEM allows differentiating the effects among variables, into direct and indirect effects. Direct effects are those where one variable directly influences another variable, without intermediate variables. Indirect effects are those where the influence of one variable on another is mediated by one or more variables. In SEM, variables are classified as either exogenous or endogenous; where exogenous variables are independent variables determined outside of the system, and endogenous variables are determined within the system and act as dependent variables.

Causal relationships among traits are present in many biological phenomena [20]; however, neither marker assisted selection (MAS) nor genomic selection procedures account for such causation. In the present research, following Smith–Hazel index theory [2,3], a selection index was developed that takes into account the causal structure among yield component traits by using coupling selection index (SI) theory and structural equation modeling (SEM). We found that when causal information is used in the index, its selection efficiency for finding better genotypes improves.

2. Materials and Methods

2.1. Path Coefficients

All parameters estimated were assumed to be normal, either due to the normality of the data or large sample sizes. A t test was used to test the null hypothesis that a path coefficient is equal to zero in the population. A path coefficient may be different from zero if its absolute value exceeds 1.96, 2.58, and 3.30 (two-tailed test) at the level of significance $p \leq 0.05$, $p \leq 0.01$, and $p \leq 0.001$, respectively [21]. Using a standardized path coefficient

helps make comparisons among them. In the present research, analysis was conducted using PROC CALIS in SAS 9.2 with the ML method.

2.2. Causal Coefficients as Economic Weights

When using SI for multiple traits simultaneously, it is common to assign an economic weight to each trait involved in the SI [22]. The vector of economic weights can be set or can be estimated. In this research we used the causal path directly and total effects estimated from the data (see Section 2.6).

2.3. Experimental Data

The data consisted of a population of two check varieties and 280 winter wheat lines evaluated at five locations in the state of Nebraska (Lincoln, Mead, McCook, Clay Center and Alliance) during the winter of 2013. The wheat lines were in the F6 generation of the winter wheat UNL program, which were selected based on the experience of the wheat breeder from around 1400 lines in the F5. The experimental design in each location was an augmented incomplete block design, with two replicated check cultivars (Goodstreak and Camelot). There were ten incomplete blocks, each of which consisted of 28 experimental lines and two check varieties. Goodstreak and Camelot are varieties that are well-adapted to different regions of Nebraska. To evaluate grain yield and yield components, we took random samples of 10 spikes per plot, including the new lines and the two checks, making a total of 300 samples per location. The agronomy data used in this research included grain yield (YLD), spikes per square meter (SPSM), kernels per spike (KPS), and kernel weight (KW). Grain yield was measured by using a combine harvester at each plot in each location. Spikes per square meter was estimated based on grain yield, kernel per spike, and kernel weight. Kernels per spike was estimated by the average of counting the number of kernels when threshing 10 spikes. Kernel weight was measured after weighing the total number of seeds of 10 spikes and dividing by the total number of seeds. When conducting a SEM, a sufficient sample size should be 100–200 samples or 5–10 times the number of parameters in the model [23]. The winter wheat data set consisted of a sample size of 1500, which exceeds the minimum recommended. The DNA marker dataset included 231 DArt markers declared significant in a previous analysis [24].

2.4. Data Analysis

To estimate the genotypic, phenotypic, and environmental variance–covariance matrices ($\hat{\Sigma}_g$, $\hat{\Sigma}_y$, $\hat{\Sigma}_e$) involved in the SI without molecular markers and the coefficient matrix (Λ), we performed a two-stage analysis. The first stage consisted of the estimation of the variance–covariance matrices using the method of moment estimators based on the sums of squares and cross-products (SSCP) matrices from the multivariate analysis of variance (MANOVA) with the linear model (1). The second stage was performed using MANOVA to fit a linear model, to remove the main effects of environment and blocks for each trait, and modeling the residuals to estimate the coefficient matrix (Λ) [13,16].

2.5. Estimation and Covariance Matrices

The linear model to estimate the variance–covariance matrices that are needed for the SI is the model for two-way crossed classification:

$$\boldsymbol{Y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{E}_i + \boldsymbol{B}(\boldsymbol{E})_{ki} + \boldsymbol{G}_j + (\boldsymbol{GE})_{ij} + \boldsymbol{\varepsilon}_{ijk} \quad (1)$$

where $\boldsymbol{Y}_{ijk}$ is the vector for the p traits of the j^{th} genotype in the i^{th} environment for the k^{th} block; $\boldsymbol{\mu}$ is the overall mean vector, $\boldsymbol{E}_i$ is the vector of main effects of the i^{th} environment; $\boldsymbol{B}(\boldsymbol{E})_{ki}$ is the vector of effects of the k^{th} block nested in the i^{th} environment; $\boldsymbol{G}_j$ is the vector of main effects of the j^{th} genotype; $\boldsymbol{GE}_{ij}$ is the vector of interaction effects between the j^{th} genotype in the i^{th} environment, and $\boldsymbol{\varepsilon}_{ijk}$ is the vector of the experimental residual for the j^{th} genotype in the i^{th} environment into the k^{th} block. Since we assumed the basic

genetic model that did not contain a GE interaction, we used $\boldsymbol{\varepsilon}^*_{ijk} = (\boldsymbol{GE})_{ij} + \boldsymbol{\varepsilon}_{ijk}$ in place of the last two terms. It was suggested that since the variation of the interaction is more environmental, it is reasonable to pool the variance of $\boldsymbol{\varepsilon}_{ijk}$ ($\Sigma_{\varepsilon_{ijk}}$) with the GE interaction variance (Σ_{gxe}) [25].

The genotype (Σ_{g}) and environmental (Σ_{e}) covariance matrices were estimated using the method of moments, based on a multivariate analysis of variance (MANOVA) for balanced data in SAS 9.2 with the procedure PROC GLM [26].

The linear model for estimating the variance–covariance matrices that include molecular markers as a random variable is:

$$\boldsymbol{Y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{E}_i + \boldsymbol{B}(\boldsymbol{E})_{ki} + \mathbf{M}_j + \mathbf{R}_j + \boldsymbol{\varepsilon}_{ijk} \tag{2}$$

where $\mathbf{M}_j$ is the vector of main effects of the j^{th} molecular marker, and $\mathbf{R}_j$ is the j^{th} vector of genetic residuals (r), which is part of the genetic variation (g) not explained by the j^{th} molecular marker (m), that is $\Sigma_{\text{g}} = \Sigma^*_{\text{m}} * \Sigma_{\text{r}}$. Each marker had values of -1, 0, and 1. The other parameters are the same as described in Equation (1), and we assumed the genetic residual and the model error were independent and normally distributed. We still assumed the basic genetic model, which did not contain an **ME** interaction. We pooled the variance of $\boldsymbol{\varepsilon}^*_{ijk} = \mathbf{R}_j + \boldsymbol{\varepsilon}_{ijk}$, ($\Sigma_{\varepsilon^*_{ijk}}$) with the environmental variance (Σ_{e}) [25] to make the phenotypic variance–covariance matrix for computing the vector of weights (**b**).

2.6. Estimation of Coefficient Matrix (Λ)

The second stage focuses on the estimation of the coefficient matrix (Λ). In this stage, grain yield and yield components were analyzed using SEM with observable variables by modeling the Y's residuals (YLD_{R}, SPSM_{R}, KPS_{R}, and KW_{R}). These residuals were obtained by subtracting the main effects of environment and blocks within the environment from the observed values [3].

$$\boldsymbol{r}_{ijk} = \boldsymbol{Y}_{ijk} - (\boldsymbol{\mu} + \boldsymbol{E}_i + \boldsymbol{B}(\boldsymbol{E})_{ki}) \tag{3}$$

The estimation of the coefficient matrix (Λ), which accounts for the causal relationship between grain yield and yield components, was conducting using PROC CALIS in SAS 9.2 software with the maximum likelihood estimation method and following the recommended steps [21]. The chi-squared (χ^2) test is the most widely used for testing the significance of the difference between sample covariances (**Σ**) and the predicted covariance ($\boldsymbol{\Sigma}(\theta)$) matrices. Along with the chi-square test, there are other model-fit criteria, such as GFI, AGF, I and NFI, that are described in the model evaluation section. Non-significant chi-square and values greater than 0.90 for GFI, AGFI, and NFI were used to evaluate the final model.

To conduct the SEM it is necessary to have prior knowledge of the causal relationships among the traits. For this research the unidirectional causal relationship between grain yield and yield components in small grains was used [8] (see Figure 1). The same estimated coefficient (Λ) matrix was used for both causal models, i.e., with and without molecular markers.

2.7. SEM Model Methodology

This research was based on the causal unidirectional relationship between grain yield and yield components in small grains [8]. Figure 1 shows the sequential development of yield components, where the later components are influenced by the earlier ones. Using structural equation modeling (SEM) it is possible to estimate the causal relationships present among grain yield and its components for small grains.

In general, the structural equation model with the observed variables can be written as:

$$Y = \mathbf{B}Y + \boldsymbol{\Gamma}X + \zeta \tag{4}$$

where Y is a $px1$ vector of endogenous variables ($p = 4$): yield (YLD), spikes per square meter (SPSM), kernels per spike (KPS), and kernel weight (KW); $\mathbf{B}$ is the pxp coefficient matrix expressing the causal relationship among endogenous variables, which is commonly a triangular matrix with zeros on its diagonal; X is the $qx1$ vector of exogenous variables; $\mathbf{\Gamma}$ is the pxq coefficient matrix expressing the causal relationship among endogenous and exogenous variables; ζ is the $px1$ disturbance vector, assumed to have $\mathrm{E}(\zeta) = 0$ and a covariance matrix $\mathrm{E}(\zeta\zeta) = \mathbf{\Psi}$ and also assumed to be uncorrelated with exogenous variables. It is also assumed that $\Lambda = (\mathrm{I} - \mathbf{B})$ is nonsingular.

For a case in which we only have endogenous variables, based on Figure 1, $\mathbf{B}$ can be

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \beta_{12} & 0 & 0 & 0 \\ \beta_{13} & \beta_{23} & 0 & 0 \\ \beta_{14} & \beta_{24} & \beta_{34} & 0 \end{pmatrix} \tag{5}$$

note that Equation (4) can be written in reduced form, as

$(\mathbf{I} - \mathbf{B})Y = \mathbf{\Gamma}X + \zeta$, let $\Lambda = (\mathbf{I} - \mathbf{B})$, $\Lambda Y = \mathbf{\Gamma}X + \zeta$, $Y = \Lambda^{-1}\mathbf{\Gamma}X + \Lambda^{-1}\zeta$, since $\Lambda^{-1} = (\mathbf{I} - \mathbf{B})^{-1}$ exists and by letting $\Pi = \Lambda^{-1}\mathbf{\Gamma}$ and $\boldsymbol{\nu} = \Lambda^{-1}\zeta$. Finally, the reduced form of the model is

$$Y = \Pi X + \boldsymbol{\nu} \tag{6}$$

Estimating parameters in SEM is unlike multiple regression and ANOVA, since the estimated parameters are obtained by minimizing the difference between the sample covariances ($\mathbf{\Sigma}$) and the predicted covariance ($\mathbf{\Sigma}(\boldsymbol{\theta})$), where $\boldsymbol{\theta}$ is the vector that contains the model parameters ($\mathbf{B}$, $\mathbf{\Gamma}$ and $\boldsymbol{\psi}$). If the model is correct, the population covariance matrix is equal to the model predicted covariances, $\mathbf{\Sigma} = \mathbf{\Sigma}(\theta)$ [19]. In this case the model implied population covariance matrix is based on Y and X, and has the order (p + q) × (p + q), where

$$\mathbf{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \tag{7}$$

Each submatrix of the model's implied covariance matrix ($\mathbf{\Sigma}(\boldsymbol{\theta})$) can be obtained as $\Sigma_{yy} = E(YY') = (\mathbf{I} - \mathbf{B})^{-1}\left(\mathbf{\Gamma}\Sigma_{xx}\mathbf{\Gamma}' + \psi\right)(\mathbf{I} - \mathbf{B})^{-1'}$, $\Sigma_{xx} = E(XX') = \mathbf{\Sigma}_{xx}$, and $\Sigma_{yx} = E(YX') = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Gamma}\Sigma_{xx}$. Finally,

$$\mathbf{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} (\mathbf{I} - \mathbf{B})^{-1}\left(\mathbf{\Gamma}\Sigma_{xx}\mathbf{\Gamma}' + \psi\right)(\mathbf{I} - \mathbf{B})^{-1'} & (\mathbf{I} - \mathbf{B})^{-1}\Gamma\Sigma_{xx} \\ \Sigma_{xx}\mathbf{\Gamma}'(\mathbf{I} - \mathbf{B})^{-1'} & \Sigma_{xx} \end{pmatrix} \tag{8}$$

The parameter vector $\boldsymbol{\theta}$ is estimated by minimizing the distance between the model's implied covariance matrix $\mathbf{\Sigma}(\boldsymbol{\theta})$ and the observed covariance matrix $\hat{\mathbf{\Sigma}}$ using a criterion of "closeness".

2.8. SEM Model Evaluation

There are several statistics used to evaluate the model fit, which consists of measuring the validity of the hypothesis that $\mathbf{\Sigma} = \mathbf{\Sigma}(\theta)$. The fundamental hypothesis for the SEM is that the matrix of covariance of the observed variables is a function of a set of parameters [19]. If the model is correct, and if the parameters are known, the population covariance matrix will be exactly reproduced. Where $\mathbf{\Sigma}$ is the population covariance, and $\mathbf{\Sigma}(\theta)$ is the covariance matrix as a function of the model parameters (θ). Some of these statistics are: chi-square statistics (χ^2), goodness of fit index (GFI), adjusted goodness of fit index (AGFI), and root mean square error of the approximation (RMSEA). The root mean square error of approximation (RMSEA) is one of the most informative fit indices, due to its sensibility to the number of estimated parameters in the model; the range of the RMSEA is between 0.05 to 0.10, where a value less or equal to 0.05 shows a good fit and values above 0.10 indicate a poor fit [27].

2.9. Model Comparison

One selection index is preferred to another if it improves selections in some sense. To compare if coupling SEM with the Smith–Hazel index improved the selection efficiency, we used two criteria:

Relative efficiency (RE) of the selection index. This is expressed as the ratio of the correlation between the selection index and the breeding value (ρ_{HI}) for two different selection indices [22].

Mean square error of prediction. This is a criterion used as a way to evaluate and compare SI's [28]. When the correlation is large between H and I, this means that the index (I) will better predict the breeding value (H); that the mean squared error of the breeding prediction will be small and the effectiveness of the prediction of the SI will be greater. The term $\sigma_H^2\left(1-\rho_{HI}^2\right)$ was called the mean square error of prediction, and the ratio $\frac{\sigma_H^2\left(1-\rho_{HI}^2\right)}{\sigma_H^2} = 1-\rho_{HI}^2$ was considered the effectiveness of I in predicting H [29]. Therefore, the SI will be more effective for predicting H when ρ_{HI}^2 is large.

2.10. Model Validation

To validate the proposed selection indices with causal structure a simulation study was conducted. The process consisted of simulating grain yield (YLD) and yield components (SPSM, KPS, and KW) with a recursive structure of a population of 282 winter wheat genotypes as fixed effects, with 5 environments, 2 replications, and residual as random effects. It is important to point out that the true population consisted of 282 wheat genotypes tested in 5 environments. The means of the 282 true genotypes for each trait for all 5 environments were standardized and used as true fixed effects for genotypes in the simulation model. The causal relationships among traits were taken into account in the simulation program by including the estimated coefficients as the true path coefficients from the true population (see Figure 1). All responses were simulated as multivariate normal random variables using the rnorm function in R. A total of 500 data sets were simulated. The validity of the simulation was assessed using the correlation between the real yield data and the simulated yield data.

3. Results

3.1. Models Developed

Coupling Causal Structure and MM into the Smith–Hazel Index

Coupling SEM to the Smith–Hazel index was developed in two scenarios. The first scenario, named model 1, is the Smith–Hazel index with causality among traits (Figure 2). The second scenario has both the molecular markers and the causality (Figure 3).

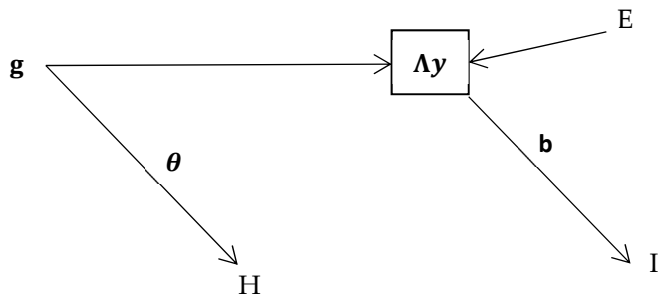

Figure 2. Path diagram of the Smith–Hazel index, in which the rectangle for **y** refers to a vector of an observable variable (the phenotypic values). The matrix **Λ** accounts for the causal relationships among traits. The vector of phenotypic values is influenced by the vector of genotypic value (**g**) and the vector of environment (E). There are also two important parts in the path, the selection index (I), and the merit or breeding value (H).

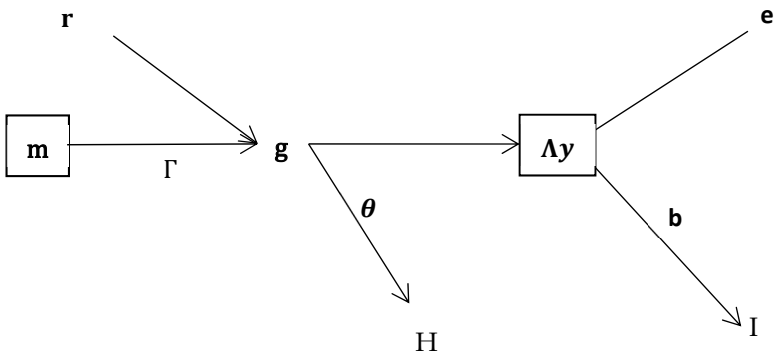

Figure 3. Path diagram showing the vector of phenotypic values (**y**) because of the effects of the latent endogenous variable of the vector of genotype (**g**) and the vector of environment (**e**). On the other hand, the vector of the latent endogenous variable (**g**) could be explained by the vector of molecular markers (**m**) and the vector of the genetic residual (**r**), which is a part of the variation not explained by markers. The rectangles refer to observable variables, while the circle refers to a latent variable (not observed).

Coupling 1: Smith–Hazel with causality among traits

The model in Figure 2 can be represented by using structural equation modeling with observed variables [10,11,30].

$$\boldsymbol{\Lambda y} = \mathbf{g} + \mathrm{E} \tag{9}$$

where **y** is a ($p \times 1$) vector of phenotypic values, **g** is the additive genotypic value, E is the vector of environmental effect, and $\boldsymbol{\Lambda}$ is a ($p \times p$) matrix of structural coefficients, accounting for the causal relationship between the p traits.

Model 9 can be written in a reduced form as:

$$\boldsymbol{y} = \Lambda^{-1}\mathbf{g} + \Lambda^{-1}\mathrm{E} \tag{10}$$

Defining the basic genetic model as $\boldsymbol{y} = \mathbf{g}^* + \mathrm{E}^*$ where $\mathbf{g}^* = \Lambda^{-1}\mathbf{g}$ and $\mathrm{E}^* = \Lambda^{-1}\mathrm{E}$, and where $\mathbf{g}^*$ and E^* are the genetic and the environmental effects after accounting for causal structures.

From Figure 2, the index is $\mathrm{I} = \mathbf{b}'\boldsymbol{y}$, and the merit is $\mathrm{H} = \boldsymbol{\theta}'\mathbf{g}^*$. Following the Smith–Hazel theory for maximizing **b**, i.e., we need to maximize the correlation between H and I (ρ_{HI}). By assuming an independence between **g** and E, we can compute $\rho_{\mathrm{HI}} = \frac{Cov(H,I)}{\sqrt{Var(\mathrm{H})Var(\mathrm{I})}}$ whose corresponding variances and covariance are:

$$Var(\mathrm{I}) = Var\left(\mathbf{b}'\boldsymbol{y}\right) = \mathbf{b}'Var(\boldsymbol{y})\mathbf{b} = \mathbf{b}'Var\left(\Lambda^{-1}\mathbf{g} + \Lambda^{-1}\mathrm{E}\right)\mathbf{b} = \mathbf{b}'\left[\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'} + \Lambda^{-1}\Sigma_{\mathrm{E}}\Lambda^{-1'}\right]\mathbf{b}$$

$$Var(\mathrm{H}) = Var\left(\boldsymbol{\theta}'\mathbf{g}^*\right) = \boldsymbol{\theta}'Var\left(\Lambda^{-1}\mathbf{g}\right)\boldsymbol{\theta} = \boldsymbol{\theta}'\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'}\boldsymbol{\theta}$$

$$Cov(H,I) = Cov\left(\boldsymbol{\theta}'\mathbf{g}^*,\ \mathbf{b}'\mathbf{y}\right) = \boldsymbol{\theta}'Cov(\mathbf{g}^*,\ \mathbf{y})\mathbf{b} = \boldsymbol{\theta}'Cov\left(\Lambda^{-1}\mathbf{g},\ \Lambda^{-1}\ \mathbf{g} + \Lambda^{-1}\mathrm{E}\right)\mathbf{b} = \boldsymbol{\theta}'\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'}\mathbf{b}$$

then, substituting the corresponding variances and covariances in ρ_{HI} we end up with the following equation

$$\rho_{\mathrm{HI}} = \frac{Cov(H,I)}{\sqrt{Var(\mathrm{H})Var(\mathrm{I})}} = \frac{\boldsymbol{\theta}'\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'}\mathbf{b}}{\sqrt{\boldsymbol{\theta}'\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'}\boldsymbol{\theta}}\sqrt{\mathbf{b}'\left[\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'} + \Lambda^{-1}\Sigma_{\mathrm{E}}\Lambda^{-1'}\right]\mathbf{b}}}$$

since $\boldsymbol{\theta}$ is a vector of constants, we only need to maximize the following equation

$$\frac{\boldsymbol{\theta}'\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'}\mathbf{b}}{\sqrt{\mathbf{b}'\left[\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'} + \Lambda^{-1}\Sigma_{\mathrm{E}}\Lambda^{-1'}\right]\mathbf{b}}} = \frac{\boldsymbol{\theta}'\Sigma^*_{\mathbf{g}}\mathbf{b}}{\sqrt{\mathbf{b}'\Sigma^*_{y}\mathbf{b}}}$$

taking derivatives of this expression with respect to **b** and set to zero, we end up with the optimal weights being expressed as

$$\Sigma_y^* \mathbf{b} = \Sigma_{\mathbf{g}}^* \theta$$

then

$$\mathbf{b} = \Sigma_y^{*-1} \Sigma_{\mathbf{g}}^* \theta = \left[\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'} + \Lambda^{-1}\Sigma_{\mathrm{E}}\Lambda^{-1'}\right]^{-1}\left[\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'}\right]\theta$$

$$\mathbf{b} = \Lambda'\left[\Sigma_{\mathrm{g}} + \Sigma_{\mathrm{E}}\right]^{-1}\Lambda\Lambda^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'}\theta \text{ since } (AB)^{-1} = B^{-1}A^{-1}$$

$$\mathbf{b} = \Lambda'\left[\Sigma_{\mathrm{g}} + \Sigma_{\mathrm{E}}\right]^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'}\theta = \Lambda'\Sigma_y^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'}\theta \tag{11}$$

Finally, the Smith–Hazel index that accounts for the causal relationship between traits is

$$\mathrm{I}_c = \mathbf{b}'\boldsymbol{y} = \left(\Lambda'\Sigma_{\boldsymbol{y}}^{-1}\Sigma_{\mathrm{g}}\Lambda^{-1'}\theta\right)'\boldsymbol{y} = \theta'\Lambda^{-1}\Sigma_{\mathrm{g}}\Sigma_y^{-1}\Lambda\,\boldsymbol{y} \tag{12}$$

where the subscript *c* stands for causality. Therefore, Equation (16) is the expression that couples SEM and the Smith–Hazel index.

Coupling 2: Smith–Hazel index with molecular markers and causality among traits

From the path diagram shown in Figure 3 we can see that the genotypic value (**g**) can be expressed as $\mathbf{g} = \Gamma\mathbf{m} + \mathbf{r}$, where **m** represents the vector of molecular markers and **r** the vector of the portion of the genetic value that it is not explained by molecular markers, and Γ is a $(p \times m)$ matrix where p is the number of traits and m is the number of markers.

The model in Figure 3, can be represented using a SEM with observed variables. The structural model can be represented as

$$\Lambda\boldsymbol{y} = \Gamma\mathbf{m} + \mathbf{r} + \mathrm{E} \tag{13}$$

where $\boldsymbol{y}$ is a 4×1 vector of endogenous variables: YLD, SPSM, KPS, and KW; $\Lambda = (\mathrm{I} - \mathbf{B})$ is the 4×4 nonsingular matrix that includes the coefficient matrix $(\mathbf{B})$ expressing the causal relationship among endogenous variables; **m** is the $(m \times 1)$ vector of molecular markers; $\boldsymbol{\Gamma}$ is the $(p \times m)$ coefficient matrix, where p is the number of traits; **r** is the (4×1) residual vector assumed to have $\mathrm{E}(\mathbf{r}) = 0$ and uncorrelated with **m**.

The reduced form of the model can be written as

$$\boldsymbol{y} = \Lambda^{-1}\Gamma\mathbf{m} + \Lambda^{-1}\mathbf{r} + \Lambda^{-1}\mathrm{E} \tag{14}$$

now define $\boldsymbol{y} = \mathbf{g}^* + \mathrm{E}^*$, where $\mathbf{g}^* = \Lambda^{-1}\Gamma\mathbf{m} + \Lambda^{-1}\mathbf{r}$ and $\mathrm{E}^* = \Lambda^{-1}\mathrm{E}$ with the index $\mathrm{I} = \mathbf{b}'\boldsymbol{y}$, and the breeding value $\mathrm{H} = \theta'\mathbf{g}^*$.

Following the Smith–Hazel theory for maximizing the vector of weights (**b**), we need to maximize the correlation between H and I. Assuming independence among **m**, **r,** and E.

$$\begin{aligned} Var(\mathrm{I}) = Var\left(\mathbf{b}'\boldsymbol{y}\right) &= \mathbf{b}'Var(\boldsymbol{y})\mathbf{b} = \mathbf{b}'Var\left(\Lambda^{-1}\Gamma\mathbf{m} + \Lambda^{-1}\mathbf{r} + \Lambda^{-1}\mathrm{E}\right)\mathbf{b} \\ &= \mathbf{b}'\left[\Lambda^{-1}\Gamma\Sigma_{\mathrm{m}}\Gamma'\Lambda^{-1'} + \Lambda^{-1}\Sigma_{\mathrm{r}}\Lambda^{-1'} + \Lambda^{-1}\Sigma_{\mathrm{E}}\Lambda^{-1'}\right]\mathbf{b} \end{aligned}$$

$$\begin{aligned} Var(\mathrm{H}) = Var\left(\theta'\mathbf{g}^*\right) &= \theta'Var\left(\Lambda^{-1}\Gamma\mathbf{m} + \Lambda^{-1}\mathbf{r}\right)\theta \\ &= \theta'\left[\Lambda^{-1}\Gamma\Sigma_{\mathrm{m}}\Gamma'\Lambda^{-1'} + \Lambda^{-1}\Sigma_{\mathrm{r}}\Lambda^{-1'}\right]\theta \end{aligned}$$

$$\begin{aligned} Cov(H, I) = Cov\left(\theta'\mathbf{g}^*,\ \mathbf{b}'\boldsymbol{y}\right) &= \theta'Cov(\mathbf{g}^*,\ \mathbf{y})\mathbf{b} \\ &= \theta'Cov\left(\Lambda^{-1}\Gamma\mathbf{m} + \Lambda^{-1}\mathbf{r},\ \Lambda^{-1}\Gamma\mathbf{m} + \Lambda^{-1}\mathbf{r} + \Lambda^{-1}\mathrm{E}\right)\mathbf{b} \\ &= \theta'\Lambda^{-1}\left[\Gamma\Sigma_{\mathrm{m}}\Gamma' + \Sigma_{\mathrm{r}}\right]\Lambda^{-1'}\mathbf{b} \end{aligned}$$

$$\rho_{HI} = \frac{Cov(H,I)}{\sqrt{Var(H)Var(I)}}$$
$$= \frac{\theta' \Lambda^{-1}\left[\Gamma\Sigma_m\Gamma' + \Sigma_r\right]\Lambda^{-1'}\mathbf{b}}{\sqrt{\theta'\left[\Lambda^{-1}\Gamma\Sigma_m\Gamma'\Lambda^{-1'} + \Lambda^{-1}\Sigma_r\Lambda^{-1'}\right]\theta}\sqrt{\mathbf{b}'\left[\Lambda^{-1}\Gamma\Sigma_m\Gamma'\Lambda^{-1'} + \Lambda^{-1}\Sigma_r\Lambda^{-1'} + \Lambda^{-1}\Sigma_E\Lambda^{-1'}\right]\mathbf{b}}}$$

since θ is a constant, we only need to maximize

$$\frac{\theta'\left[\Lambda^{-1}\Gamma\Sigma_m\Gamma'\Lambda^{-1'} + \Lambda^{-1}\Sigma_r\Lambda^{-1'}\right]\mathbf{b}}{\sqrt{\mathbf{b}'\left[\Lambda^{-1}\Gamma\Sigma_m\Gamma'\Lambda^{-1'} + \Lambda^{-1}\Sigma_r\Lambda^{-1'} + \Lambda^{-1}\Sigma_E\Lambda^{-1'}\right]\mathbf{b}}} = \frac{\theta'\Sigma_{\mathbf{m}}^{***}\mathbf{b}}{\sqrt{\mathbf{b}'\Sigma_y^{***}\mathbf{b}}}$$

taking derivatives this expression with respects to **b** and set equal to zero, we end up with the optimal weights expressed as

$$\Sigma_y^{***}\mathbf{b} = \Sigma_{\mathbf{m}}^{***}\theta$$

Then

$$\mathbf{b} = \Sigma_y^{***-1}\Sigma_{\mathbf{m}}^{***}\theta$$

$$\mathbf{b} = \left[\Lambda^{-1}\Gamma\Sigma_m\Gamma'\Lambda^{-1'} + \Lambda^{-1}\Sigma_r\Lambda^{-1'} + \Lambda^{-1}\Sigma_E\Lambda^{-1'}\right]^{-1}\left[\Lambda^{-1}\Gamma\Sigma_m\Gamma'\Lambda^{-1'} + \Lambda^{-1}\Sigma_r\Lambda^{-1'}\right]'\theta$$

$$\mathbf{b} = \Lambda'\left[\Gamma\Sigma_m\Gamma' + \Sigma_r + \Sigma_E\right]^{-1}\Lambda\Lambda^{-1}\left[\Gamma\Sigma_m\Gamma' + \Sigma_r\right]\Lambda^{-1'}\theta$$

$$\mathbf{b} = \Lambda'\left[\Gamma\Sigma_m\Gamma' + \Sigma_r + \Sigma_E\right]^{-1}\left[\Gamma\Sigma_m\Gamma' + \Sigma_r\right]\Lambda^{-1'}\theta \quad (15)$$

Finally, the Smith–Hazel index with molecular markers and causality among the traits is

$$I = \mathbf{b}'\boldsymbol{y} = \left(\Lambda'\left[\Gamma\Sigma_m\Gamma' + \Sigma_r + \Sigma_E\right]^{-1}\left[\Gamma\Sigma_m\Gamma' + \Sigma_r\right]\Lambda^{-1'}\theta\right)'\boldsymbol{y}$$

$$I = \theta'\Lambda^{-1}\left[\Gamma\Sigma_m\Gamma' + \Sigma_r\right]\left[\Gamma\Sigma_m\Gamma' + \Sigma_r + \Sigma_E\right]^{-1'}\Lambda\,\boldsymbol{y} \quad (16)$$

$$I = \theta'\Lambda^{-1}[\Sigma_m^* + \Sigma_r][\Sigma_m^* + \Sigma_r + \Sigma_E]^{-1'}\Lambda\,\boldsymbol{y} \quad (17)$$

Table 1 shows the models developed considering only causality and both causality and molecular markers.

Table 1. Different expressions of the selection index considering molecular markers and causality among traits.

Model	Selection Index (I)
Classical Smith-Hazel	$I = \mathbf{b}'\boldsymbol{y} = \theta'\Sigma_g\Sigma_y^{-1}\boldsymbol{y}$
Smith-Hazel with causality	$I = \mathbf{b}'\boldsymbol{y} = \theta'\Lambda^{-1}\Sigma_g\left[\Sigma_g + \Sigma_E\right]^{-1'}\Lambda\,\boldsymbol{y}$
Smith-Hazel with markers and causality	$I = \mathbf{b}'\boldsymbol{y} = \theta'\Lambda^{-1}[\Sigma_m^* + \Sigma_r][\Sigma_m^* + \Sigma_r + \Sigma_E]^{-1'}\Lambda\,\boldsymbol{y}$

Figure 4 and Table 2 display the path coefficients from the final model, which fit well since the model is exactly identified; i.e., the total number of parameters estimated (t = 10, 6 path coefficients and 4 errors) is equal to the number of data points $\left(\frac{p(p+1)}{2} = 10\right)$. The six coefficients were all significant, $p \leq 0.05$. Since standardized data were used in the analysis, direct comparisons among grain yield and yield components coefficients were possible, and it was easy to understand how they impacted the yield. For example, the direct positive effect of SPSM on yield (0.81) was greater than the indirect effect on KW (−0.347), which means that the net effect of increasing SPSM would be to increase YLD. Note that, the indirect effects are calculated by multiplying the path coefficients of each path of the associated variable to the dependent variable. For instance, the indirect effect of KPS on YLD follows the path KPS→KW→YLD is (−0.25) (0.17) = −0.043. The total effect

is just the sum of direct and indirect effects. In the case of SPSM, which has three indirect paths, the indirect effect is the summation of the indirect effects. Spikes per square meter showed the biggest effect (0.81) on grain yield, followed by kernel per spike (0.52) and kernel weight (0.17). That is, the effect of increasing SPSM would increase YLD more than if we increased the effect of KW on YLD.

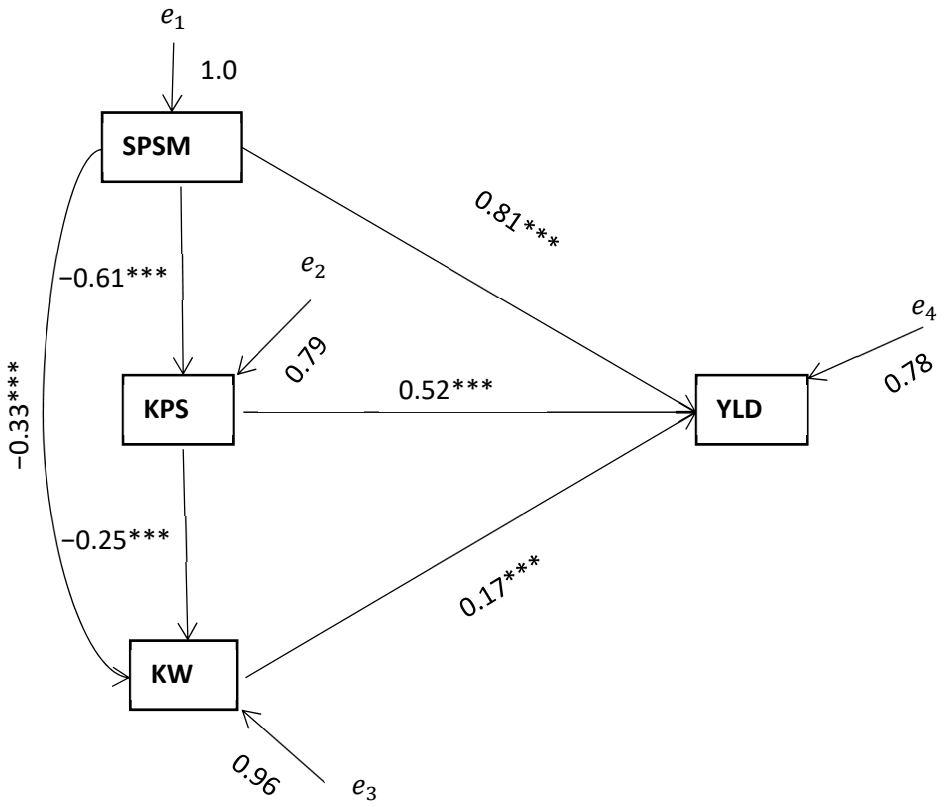

Figure 4. Standardized estimates of path coefficients of a structural equation model of grain yield (YLD) and yield components (spikes per square meter (SPSM), kernels per spike (KPS), and kernel weight (KW)). Arrows represent the direction of the variables' influence, and the numbers on the arrow lines represent the estimated standardized coefficients. Significant level: ***$p \leq 0.001$.

Table 2. Direct, indirect, and total effects of yield components on grain yield.

Variables	Direct Effect (θ_D)	Indirect Effect (θ_I)	Total Effect (θ_T)
Spikes per square meter	0.81	−0.347	0.463
Kernel per spike	0.52	−0.043	0.477
Kernel weight	0.17	0	0.170

An important tool for model comparison is the relative efficiency (RE), which increased when causality was considered over the same SI without causality. For example, the RE for comparing S-H with and without causality was $RE = \frac{SI\ in\ the\ column}{SI\ in\ the\ row} = \frac{S-H\ causality}{S-H\ classic} = 120$, this means that S–H causality is 20% more efficient than S–H classic for predicting the breeding value (BV). Moreover,$RE = \frac{SI\ in\ the\ column}{SI\ in\ the\ row} = \frac{S-H\ classic}{S-H\ causality} = 84$, which means that S–H classic is less efficient than S–H causality for predicting the breeding value (BV).

Table 3 shows the best ten percent genotypes under different selection indices. One was based on yield, while the other rankings were based on the different selection indices. Note that the S–H index with causality was one of the closest to the yield selection, this is possible when total causal information is used as the economic weights for the index.

Table 3. The best 28 genotypes (Top 10%) based on different indices where the vector of economic weights (θ) was based on total causal effects.

	Yield	S-H	S-H c	S-H mc
Standardized Mean	1.62	0.115	0.866	0.592
True Mean (g/m^2) **	480.6	432.8	456.7	448.0

** The mean of the best 28 genotypes. Abbreviations: S-H = Smith-Hazel, c = causality, and mc = markers and causality.

The validity of the simulation was assessed through the correlation between the real yield data and the simulated yield data. The correlation over the 500 simulated data sets was 0.94, indicating the simulation accurately captured the structure of the real data. To validate the proposed selection index, we performed four kinds of analyses for the simulated data, comparing with the selection index without causal structure:

(1) We computed the relative efficiency (RE) for model comparison between the indices; the results showed that the S–H causality increased by 12% over S–H classic.

(2) We examined how many of the true best 28 (Top 10% based on yield) genotypes based on true indices were identified under the simulation; the result showed a 39% coincidence with respect to the genotypes selected based on yield, while with S–H classic this was 18%.

In general, the results showed that when a causal relationship among traits is accounted for, the SI and the RE increased, and the standardized mean yield of the best 10% genotypes increased with respect to the SI without causality.

4. Discussion

We found that the direct effects followed the same pattern reported by other researchers [8,13,16]; in which SPSM had the biggest effect on yield, followed by KPS and KW (Table 2). On the other hand, the indirect effects were all negative among yield components (see Figure 1 and Table 2). The correlations between the SI (I) and the breeding value (H) increased when we considered the causal relationship between grain yield and yield components in the case of the Smith–Hazel indices. For example, the correlation for the S–H causality index (0.47) was larger than the S–H classic index (0.39). When the correlation between the SI (I) and the breeding value (H) is large, then the SI will be more effective for predicting H [29].

Since yield and yield components have a causal mechanism underlying the biological processes, it may be reasonable to use these causal coefficients as economic weights [31,32]. The result of using causal path effects as economic weights improves the ability of the Sis, increasing the mean of the selected genotypes with respect to using the same economic weight for each trait. This result confirms that taking the causal biological relationship among certain traits into account can help select promising genotypes.

The idea of using causal information among traits with SI has been suggested by other authors [11,30–32]. Direct causal effects as economic weights were used for improving grain shape and grain yield in rice [31,32]. The authors concluded that using direct path coefficients as economic weights for secondary traits and the economic weight of one for the primary trait was an effective criterion of selection for improving primary traits. Similarly, we used as a vector of economic weights the total causal effects, which not only accounts for direct path coefficients but also for indirect effects. Using total causal effects as a vector of economic weights, we showed that the S–H with causality index increased the mean yield of the best 28 genotypes, and the number of matches between this index and yield per se, more than the other SIs.

Ignoring the causal association among traits leads to a loss of valuable information [8,12,13,15,16,18]. When causal structure among yield and yield components is captured in the index, the correlation improved between the index and the BV and the RE. In addition, it is important to point out that when comparing two indices without and with the causal structure among traits, the index that accounts for causal relationship increased

the mean for yield of the selected genotypes. Accounting for causality in the index could be used to help breeders with the selection of promising genotypes, because this method of selection considers the recursive relationship among traits.

5. Conclusions

The specific purpose of this research was to improve the ability of the selection index for identifying promising genotypes by accounting for the causal structure among yield and yield components in winter wheat. The results from the true data showed that the most important findings were: First, the causal Smith–Hazel indices improved the relative efficiency with respect to those that did not incorporate causal information. Second, the selection indices may help when selecting for a primary trait. Third, using total causal effects as a vector of economic weights we showed that the S–H with causality index increased the mean yield of the best 28 genotypes, and the number of matches between this index and yield per se more, than the other SIs.

The proposed multi-trait selection indices can account for the causal structure among traits when there is a prior knowledge of the causal relationships. These indices provide certain advantages over the classic Smith–Hazel index by improving the correlation between the index and the breeding value, the relative efficiency, and the mean-yield of the selected genotypes when selection is based on a primary trait. In addition, the contributions of yield components to grain yield when the selection index accounts for the causal mechanism can be seen through the increase in the mean yield.

The results indicated that among the evaluated indices, the S–H causality index is recommended for improving yield when no marker information is available. In addition, selecting for a primary trait using total causal effects as economic weights for yield contributors and one for the trait of interest should serve as an effective selection criterion for improving grain yield.

Author Contributions: Conceptualization, J.V.H.-C. and K.M.E.; Resources, S.P.B.; Writing—original draft, J.V.H.-C.; Writing—review & editing, J.V.H.-C., J.S.-R., K.M.E. and S.P.B. All authors have read and agreed to the published version of the manuscript.

Funding: Mexico's National Science and Technology Council (CONACYT-Mexico) funded this research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hazel, L.N.; Lush, J.L. The efficiency of three methods of selection. *J. Hered.* **1942**, *33*, 393–399. [CrossRef]
2. Smith, H.F. A discriminant function for plant selection. *Ann. Eugen.* **1936**, *7*, 240–250. [CrossRef]
3. Hazel, L.N. The genetic basis for constructing selection indexes. *Genetics* **1943**, *28*, 476–490. [CrossRef] [PubMed]
4. Bernardo, R. *Breeding for Quantitative Traits in Plants*, 2nd ed.; Stemma Press: Woodbury, MN, USA, 2010.
5. Neimann-Sorensen, A.; Robertson, A. The association between blood groups and several production characteristics in three Danish cattle breeds. *Acta Agric. Scand.* **1961**, *11*, 163–196. [CrossRef]
6. Smith, C. Improvement of metric traits through specific genetic loci. *Anim. Sci.* **1967**, *9*, 349–358. [CrossRef]
7. Lande, R.; Thompson, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **1990**, *124*, 743–756. [CrossRef]
8. Dofing, S.M.; Knight, C.W. Alternative model for path analysis of small-grain yield. *Crop. Sci.* **1992**, *32*, 487–489. [CrossRef]
9. Wright, S. Correlation and causation. *J. Agric. Res.* **1921**, *201*, 557–585. Available online: https://www.google.com.mx/search?q=Wright%2C+S.%2C+1921+Correlation+and+causation.+J.+Agric.+Res.+201%3A+557%E2%80%93585&source=hp&ei=afarYPGPCsfatQXdxp7IDA&iflsig=AINFCbYAAAAAYKwEec1T6njhrEiKpZoxWx3xdCpkl817&oq=Wright%2C+S.%2C+1921+Correlation+and+causation.+J.+Agric.+Res.+201%3A+557%E2%80%93585&gs_lcp=Cgdnd3Mtd2l6EANQmAxYmAxg_RZoAHAAeACAAaUBiAGlAZIBAzAuMZgBAKABAqABAaoBB2d3cy13aXo&sclient=gws-wiz&ved=0ahUKEwix6LiF_-LwAhVHba0KHV2jB8kQ4dUDCAc&uact=5 (accessed on 24 May 2021).

10. Rosa, G.J.; Valente, B.D.; de los Campos, G.; Wu, X.-L.; Gianola, D.; Silva, M.A. Inferring causal phenotype networks using structural equation models. *Genet. Sel. Evol.* **2011**, *43*, 6. [CrossRef]
11. Gianola, D.; Sorensen, D. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics.* **2004**, *167*, 1407–1424. [CrossRef]
12. Vargas, M.; Crossa, J.; Reynolds, M.P.; Dhungana, P.; Eskridge, K.M. Structural equation modelling for studying genotype environment interactions of physiological traits affecting yield in wheat. *J. Agric. Sci.* **2007**, *145*, 151. [CrossRef]
13. Dhungana, P.; Eskridge, K.M.; Baenziger, P.S.; Campbell, B.T.; Gill, K.S.; Dweikat, I. Analysis of genotype-by-environment interaction in wheat using a structural equation model and chromosome substitution lines. *Crop. Sci.* **2007**, *47*, 477–484. [CrossRef]
14. Guillen-Portal, F.R.; Stougaard, R.N.; Xue, Q.; Eskridge, K.M. Compensatory mechanisms associated with the effect of spring wheat seed size on wild oat competition. *Crop. Sci.* **2006**, *46*, 935–945. [CrossRef]
15. Kozak, M.; Bocianowski, J.; Rybinski, W. Selection of promising genotypes based on path and cluster analyses. *J. Agric. Sci.* **2008**, *146*, 85. [CrossRef]
16. Dhungana, P. *Structural Equation Modeling of Genotype x Environment Interaction*; The University of Nebraska-Lincoln: Lincoln, NE, USA, 2004.
17. Lamb, E.; Shirtliffe, S.; May, W. Structural equation modeling in the plant sciences: An example using yield components in oat. *Can. J. Plant. Sci.* **2011**, *91*, 603–619. [CrossRef]
18. Heineck, G.C.; Ehlke, N.J.; Altendorf, K.R.; Denison, R.F.; Jungers, J.M.; Lamb, E.G.; Watkins, E. Relationships and influence of yield components on spaced-plant and sward seed yield in perennial ryegrass. *Grass Forage Sci.* **2020**, *75*, 424–437. [CrossRef]
19. Bollen, K.A. *Structural Equations with Latent Variables*; Wiley: New York, NY, USA, 1989. Available online: https://www.wiley.com/en-ae/Structural+Equations+with+Latent+Variables-p-9780471011712 (accessed on 21 May 2021).
20. Shipley, B. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference with R*; Cambridge University Press: Cambridge, UK, 2016.
21. Hatcher, L. *Norm O'Rourke. A Step-By-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling*; SAS Institute: Caray, NC, USA, 2013.
22. Baker, R.J. *Selection Indices in Plant Breeding*; CRC Press Inc.: Boca Raton, FL, USA, 1986; 218p.
23. Grace, J.B. *Structural Equation Modeling and Natural Systems*; Cambridge University Press: Cambridge, UK, 2006.
24. El-basyoni, I.S. *Association Mapping for Important Biotic & Abiotic Related Traits in Structured Wheat Population*; ETD collection for University of Nebraska—Lincoln: Lincoln, NE, USA, 2012; p. AAI3508512.
25. Falconer, D.S.; Mackay, T.P.P. *Introduction to Quantitative Genetics*; Longmans: Harlow, Essex, UK, 1996.
26. SAS Institute. *SAS/STAT 9.2 User's Guide*; SAS Institute: Caray, NC, USA, 2000.
27. MacCallum, R.C.; Browne, M.W.; Sugawara, H.M. Power analysis and determination of sample size for covariance structure modeling. *Psychol. Methods* **1996**, *1*, 130. [CrossRef]
28. Lange, C.; Whittaker, J.C. On prediction of genetic values in marker-assisted selection. *Genetics* **2001**, *159*, 1375–1381. [CrossRef]
29. Anderson, T.W.; Anderson, T.W.; Anderson, T.W.; Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, 3rd ed.; John-Wiley & Sons: Englewood Cliffs, NJ, USA, 2003.
30. Valente, B.D.; Rosa, G.J.; Gianola, D.; Wu, X.-L.; Weigel, K. Is structural equation modeling advantageous for the genetic improvement of multiple traits? *Genetics* **2013**, *194*, 561–572. [CrossRef] [PubMed]
31. Rabiei, B.; Valizadeh, M.; Ghareyazie, B.; Moghaddam, M. Evaluation of selection indices for improving rice grain shape. *Field Crop. Res.* **2004**, *89*, 359–367. [CrossRef]
32. Sabouri, H.; Rabiei, B.; Fazlalipour, M. Use of selection indices based on multivariate analysis for improving grain yield in rice. *Rice Sci.* **2008**, *15*, 303–310. [CrossRef]