Discussion Paper

The role of causal modelling in official statistics

Frank P. Pijpers, Iris A.M. Boeters

June 2023

Contents

The primary task of national statistical institutes (NSI's) is generally accepted to be the collection of data of direct social and economic interest, for the purpose of reporting on various aggregated measures of such data. One could classify this approach as purely descriptive or observational: an answer to ”what is the current value/number/trend of $\langle item \rangle$?”. Users of the output of NSI's - policy makers, scientists, the general public - often have questions not only about the ”what” but also the ”why”, which implies exploring the potential causal connections between various indicators or their longitudinal behaviour. In a setting where controlled experiments are impossible, there are limits to the scope for such causal research. Therefore NSI's are normally reticent or equivocal concerning causality in their publications. Nevertheless, advances in causal structural modelling and 'natural experiments', as also recognised in the recent awarding of the Nobel memorial prize in economic sciences to Angrist and Imbens for their methodological contributions to the analysis of causal relationships, mean that it is timely to re-examine the potential of such techniques within the context of official statistics.

This paper is intended to explore to what extent causality research has a role in innovative output of Statistics Netherlands. In the more appropriate, cautious, approach, the potential of using data to *exclude* putative causal mechanisms appears to be the most promising, and is an active line of research. A limited proof-of-concept study on secondary school selection procedures is explored as an example of the technique but should not be interpreted as exhaustive research on counterfactual causality or counterfactual fairness within school level selection in the Netherlands.

1 Introduction

The primary task of national statistical institutes is generally accepted to be the collection of data of direct social and economic interest, and reporting on various aggregated measures of such data. In order to enable intercomparison between countries, it is important to establish common standards and definitions for the collection of data and the various processing steps required to produce the aggregates. When this concerns longitudinal data, in addition it matters to not change standards and definitions over time, to the extent that this is feasible. Operational experience shows that these two requirements, intercomparability between countries and over time, can be incompatible and there are also other reasons why long consistent and high-quality time series are difficult to collect. In particular in the context of determining causal mechanisms this is problematic. The time ordering of events, i.e. *causes* preceding *consequences*, can be embedded as a relatively weak signal in time series which might be extracted through a cross-correlation of two or more of such time series. Anomalies in those time series, due to extraneous breaks - definition changes, changes in data collection procedures etc. - can destroy such correlations and worse, create spurious apparent correlations which can even reverse the inferred causal directions from what the true situation is.

It is important therefore to investigate hypothesized causal mechanisms in various independent ways, i.e. using not just longitudinal data and rely on the time ordering, but also use cross-sectional studies using for instance counter-factuals in such data to exclude certain pathways or directions in a causal diagram. Given the importance, it is useful to briefly describe some basics of structural causal modelling (SCM) as part of this introduction, in addition to a brief discussion of what is known as predictive or Granger causality. Granger-causality and cointegration is originally presented in the papers of Granger (1969); Engle and Granger (1987). Much more background on SCM can be found in the book of Pearl (2009) and also the book Imbens and Rubin (2015) provides the framework for experimental/ observational settings and examples. In the rest of this paper a particular case study is shown, where causal reasoning is used to improve, at a fundamental level, machine learning methods which would otherwise suffer from biases already present in the data used to train such methods. This case study (section 2) is part of the MSc thesis of one of the authors of this paper (I.A.M. Boeters).

1.1 predictive causality and cointegration

In the paper of Granger (1969) the approach is presented for determining whether two regularly sampled time series X, Y are related in the sense that time series Y lags a time series X, by an integer number of the regular sampling intervals, so that the behaviour of X can be used as a predictor for Y. A generalization to 3 or more time series is straightforward. In the notation of Granger (1969), for the case of 2 variables, the focus is on *simple causal models*:

$$\begin{aligned} X_i &= \sum_{j=1}^{i} a_j X_{i-j} + \sum_{j=1}^{i} b_j Y_{i-j} + \varepsilon_i \\ Y_i &= \sum_{j=1}^{i} c_j X_{i-j} + \sum_{j=1}^{i} d_j Y_{i-j} + \eta_i \end{aligned} \tag{1.1}$$

The summation index j runs from 1, rather than 0, implying that there is no 'instantaneous causality' of X to Y or vice versa, i.e. no dependency for the same time index i. This is what the 'simple' in simple causal models refers to. The upper limit for j is set to i meaning that models are excluded where there is significant direct correlation over timescales that are longer than the entire temporal extent of the measured time series, where by definition the measurements start at index 0. Evidently the direct causation can be much more limited in temporal extent than the full length of the measured series, which in terms of the coefficients $\{a, b, c, d\}_j$ implies that there exists an m such that the coefficients with index j are $= 0$ $\forall j$ with $m < j \leq i$. The a and d coefficients express auto-correlation in the time series X and Y respectively, whereas the b and c express cross-correlations between X and Y. In the absence of mutual causation (i.e. causal loops or feedback) either the b or the c (but not both) are all $= 0$. The terms ε, η in eq. (1.1) express measurement error and other stochastic perturbations in X and Y which are presumed independent, and with finite variance (e.g. Gaussian) at all times, so that the expectation values $E[]$ satisfy:

$$\begin{aligned} E[\varepsilon_i] &= E[\eta_i] = 0 \\ E[\varepsilon_i \eta_j] &= 0 \\ E[\varepsilon_i \varepsilon_j] &= \sigma^2_{\varepsilon\, i} \delta_{ij} \\ E[\eta_i \eta_j] &= \sigma^2_{\eta\, i} \delta_{ij} \end{aligned} \tag{1.2}$$

where δ_{ij} is the Kronecker δ. If the errors are homoskedastic then the time index on the variances σ^2 can be omitted, which is done here: it corresponds to the assumption of white noise mentioned in Granger (1969).

In addition to this, Fourier transforms (FTs, or Cramer representations in the terminology of Granger (1969)) are required to complete the expressions for quantifying predictive causality. In general the Fourier transform $F(\omega)$ of a time series $f(t)$ is defined by:

$$F(\omega) = \int\limits_{-\infty}^{\infty} e^{i\omega t} f(t) \mathrm{d}t \tag{1.3}$$

which is invertible so that also:

$$f(t) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} e^{-i\omega t} F(\omega) \mathrm{d}\omega \tag{1.4}$$

If the function $f(t)$ is only known at N discretely sampled times t_i $i \in \{0, .., (N-1)\}$ which are regularly spaced, as is the case considered in Granger (1969), a discrete version of the Fourier transform can be defined as well, where the integral is replaced by summation with appropriate weights for each of the samples $f(t)$, as follows:

$$F(\omega_l) = \sum_{k=0}^{N} e^{i\omega_l t_k} f(t_k) \tag{1.5}$$

where the only values of the frequency ω_l that are necessary and sufficient to fully specify the transform F are $\omega_l = 2\pi l/T$ in which T is the total temporal extent of the measured time series and the discrete index l runs from $-N/2$ to $N/2$. This discrete FT also is invertible, just as the general expression is. It is most usual to use values of $N = 2^n$ for some n because there are computational algorithms for this case that are particularly efficient (fast Fourier transforms, FFTs), which means that $N/2$ is integer. The maximum frequency is referred to as the Nyquist frequency: $\omega_{Nyq} = \pi N/T$. Note that in fact $F(-\omega_{Nyq}) = F(\omega_{Nyq})$ and also that for $f(t)$ real-valued it holds that $F(-\omega) = F(\omega)^*$ where the superscript * signifies taking the complex conjugate.

The approach of Granger (1969) is to look at properties of the cross-correlation of X and Y in the Fourier domain and perform a test on the amplitude of the FT of the cross-correlation function. The details of how to test for the direction and magnitude of the predictive causality are in that paper. It may be helpful for insight to take a slightly different route, based on Pijpers and Wanders (1994); Pijpers (2021). For this the equations (1.1) are revisited but in a continuous form, so that if X causes Y then:

$$Y(t) = \int K_{\overrightarrow{XY}}(\tau) X(t-\tau) \mathrm{d}\tau \tag{1.6}$$

Where τ is the 'lag' or 'delay' between X and Y, and the kernel $K_{\overrightarrow{XY}}$, in some fields referred to as transfer function, expresses how strongly X feeds through into Y over an entire range of time-delays, specified by the kernel K. Just as in eq. (1.1) the relationship between X and Y is presumed to be linear. The kernel $K_{\overrightarrow{XY}}$ plays the same role as the coefficients b, and there is an equivalent kernel $K_{\overrightarrow{YX}}$ which has the same role as c in eq. (1.1), by reversing the roles of X and Y respectively. By allowing that X and Y could be the same in eq. (1.6), also the cases of non-zero coefficients a or d is covered. Note that this then does require a little more care in terms of the error treatment since then the errors cannot be completely i.i.d. In principle the formalism can be extended to non-linear relationships $G(X)$, as long as G is known and invertible. An example of the latter would be a sigmoid function:

$$\begin{aligned} G(X) &= \frac{e^{\alpha X}}{1 + e^{\alpha X}} \\ X &= \frac{1}{\alpha} \ln \left[\frac{G}{1-G} \right] \end{aligned} \tag{1.7}$$

By examining eq. (1.6) it becomes clear that the purpose of a test for causality is to establish whether $K \equiv 0$ for $\tau > 0$ in which case X does not have predictive causality for Y, or $K \equiv 0$ for $\tau < 0$ in which case Y does not have predictive causality for X. If the results of such a test are consistent with $K = 0$ everywhere there is no information transfer (predictive power) either way, i.e. no causal connection in the sense of Granger (1969). If the result were to be that there are both values for $\tau > 0$ and $\tau < 0$ for which $K \neq 0$, then either there is mutual information transfer between X and Y or more complex causality, possibly with a latent (unmeasured) factor or factors Z influencing both X and Y.

In the paper Pijpers and Wanders (1994) it is described how to determine the transfer function K in (1.6) as a function of delay time, rather than merely performing the statistical test whether or not it is equal to 0, even when the sampling is not perfectly regular. In the paper Pijpers (2021) the same problem is solved but where $X = Y$ in (1.6). In any case, this problem belongs to the class of *inverse problems*, which implies that it is *ill-posed*: there is no unique solution for K i.e. many different K would produce identical X and Y. In particular in the presence of stochastic perturbations (measurement errors) this is an unfortunate property because K tends to be extremely sensitive to errors, and normally some form of regularization is required. Note that if the relationship between X and Y is non-linear, through some link function G, there is a two step process whereby first G is determined through solving the inverse problem, and then X is determined from G by the inverse operation. However, considering that the sensitivity to stochastic perturbations of the first step is already very high, the second step may well exacerbate the problem. For instance, in the example (1.7) for G, if the first step of the inverse problem results in values of G close to 0 or close to 1 with a finite error then this may well result in values of X with an infinite margin of error: i.e. confidence intervals for X inferred to be $(-\infty, x_+)$ or (x_-, ∞) respectively.

The FT of convolution integrals such as (1.6) can be performed very simply in the Fourier domain. The FT of a convolution of two functions is the product of the FT's of those two functions. A common shorthand notation for a convolution operation is a $*$, so that:

$$FT(f * g) = FT(f)FT(g) \tag{1.8}$$

When applying this theorem (1.8) to equation (1.6), the result is:

$$FT(Y) = FT(K_{\overrightarrow{XY}})FT(X) \tag{1.9}$$

Merely dividing $FT(Y)$ by $FT(X)$ and then performing an inverse Fourier transform is possible only in the very exceptional case that there is **no** value for ω for which $FT(X) = 0$. This is another way of expressing the ill-posedness of the problem, and one standard way to deal with this, for regularly sampled series, is to apply Wiener filtering which means that $FT(K_{\overrightarrow{XY}})$ is determined through:

$$FT(K_{\overrightarrow{XY}}) = \frac{FT(Y)FT^*(X)}{FT(X)FT^*(X) + \mu\,[FT(X)FT^*(X)]_{\omega=0}} \tag{1.10}$$

where a star superscript indicates taking the complex conjugate. The second term in the denominator on the righthand side, with $\mu > 0$ is the regularization term with which the adverse effects of noise and zero's in the $FT(X)$ are suppressed. It is clear that reversing the roles of X and Y, meaning that Y is hypothesized to have predictive causality for X rather than vice versa, produces a complex conjugate in the numerator. The denominator is real-valued regardless of whether Y or X appears. If the power spectra of time series X and time series Y are similar, i.e $FT(X)FT^*(X) \approx cFT(Y)FT^*(Y)$ for some constant c, then this implies that apart from taking complex conjugates the FT of the transfer function is the same and therefore the transfer function itself is time reversed but otherwise unaffected. Even if that approximate similarity of the power spectra of X and Y does not hold, it is clear that this affects the shape of the kernel but not the overall time reversal. Causality in either direction can therefore be determined. In this way the Granger causality test can be seen as the test whether K is $\neq 0$ anywhere for either $\tau > 0$ or $\tau < 0$.

The choice of weighting parameter μ will require some care. It will tend to be the case that if the measured data X and Y improve in signal-to-noise ratio, i.e. $\sigma_\varepsilon^2 / \left[FT(X)FT^*(X)\right]_{\omega=0} \downarrow$ and / or $\sigma_\eta^2 / \left[FT(Y)FT^*(Y)\right]_{\omega=0} \downarrow$, then μ can be set to a smaller value. As noted above, it will not usually be possible to set $\mu = 0$ because even in the absence of any measurement noise, there could still be ω values for which $FT(X) = 0$ or $FT(Y) = 0$ which the regularization compensates for. Since the norm of the functions X and Y, and their Fourier transforms, remain finite, as $\omega \to \infty$ the complex modulus of $FT(X)$ and $FT(Y)$ must decrease and tend to 0. This is important because it can be seen that with $\mu > 0$ the regularization term must then always dominate in the denominator for sufficiently high ω. This means that high frequencies are suppressed in the $FT(K_{\overrightarrow{XY}})$ so sharp features and edges in $K_{\overrightarrow{XY}}(\tau)$ cannot be reconstructed properly, which is a fundamental property of inverse problems. Even if the transfer function K in reality were to have a very sharp edge at $\tau = 0$ such a feature would be impossible to reconstruct perfectly from data. In practice therefore, the requirents for the measured time series in terms of numbers of samples in the series with high singal-to-noise ratio are quite strict in order to apply this technique.

The extension to multiple variables X (jointly) causing Y, or more precisely that jointly have a predictive causality for Y, is explored in the paper Engle and Granger (1987). Before the formalism above can be applied it is helpful to consider that among the separate X there can in principle also be a predictive causal connection, which complicates determining any connection with Y. In Engle and Granger (1987) this is referred to as the co-integration of the X. The solution for this is to first perform a linear transformation on the vector of time series X, such that the resulting set $\widetilde{X}$ are mutually orthogonal in the sense that the integral over time of the product of any two elements $\widetilde{X}$ becomes $= 0$ This orthogonality ensures that the causality analysis can formally be done on each pair $\left(\widetilde{X}, Y\right)$ separately.

It should be noted, as is done in Granger (1969), that such predictive causality between X and Y does not necessarily imply that the phenomenon which is measured by X causes whatever phenomenon Y measures. There could well be an unknown or unobserved common cause Z which affects both X and Y but with different lags. In the absence of explicit knowledge of such a Z, it is often a pragmatic choice to treat X and Y as directly causally connected until proven otherwise. For what follows in this paper it is assumed that preparatory research is carried out to identify potential causal mechanisms, and therefore to be as complete as possible in identifying relevant time series, and assume that this is sufficient for the subsequent analysis. One good way to conceptualise setting up such analyses is structural causal modelling; in practice this is a way to systematise knowledge about all relevant factors that might directly or indirectly influence a certain outcome. This is the subject of the next section.

1.2 structural causal modelling

Graph theory serves as the mathematical language for causal structures and holds some basic rules and definitions. Formally, a graph is a tuple $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of *nodes* and

$\mathcal{E} \subseteq v \times v$ is the set of *edges* representing the relationships between nodes (e.g., causal relationships). A set pa_v of parents of node v is the set of source nodes of incoming edges to v, node v being the *child* and the set of *ancestors* is obtained by collecting parent nodes recursively, node v being the *descendant*. Cyclical structure in a graph considerably complicates causal inference. For this reason it is often assumed that the graph representing the causal structure is a directed acyclic graph (DAG), with edges $\mathcal{E}$ that have a certain orientation and without cyclic structures within the graph, and this assumption is also adopted in this paper. For instance see a simple DAG as in Figure 1.1 with *nodes* {X, Y, Z} and *edges* as arrows between the variables that represent a direct causal dependence. In a causal structure, missing edges are as important as the present ones.

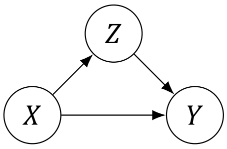


Figure 1.1 A directed acyclic graph (DAG) with *nodes* $\mathcal{V}$ {X, Y, Z} and the *directed edges* $\mathcal{E}$ as arrows between them.

A structural causal model (SCM) captures the story behind the problem and describes how the variables in a model are set. Formally, a SCM is defined as an ordered triple $\langle U, V, F \rangle$ where U denotes the set of *exogenous* variables, V denotes the set of *endogenous* variables and F denotes the set of functions mapping U to V. The *exogenous* U variables have no *parents*, which means the causes of these factors are not explained by the model and they represent unknown or random variability in the model. The *endogenous* V variables have at least one *parent*, by which the corresponding mathematical function describes how the variable state is calculated. For instance, consider the following SCM (see also fig. 1.1):

$$
\begin{aligned}
U = \{U_X, U_Y, U_Z\}, \qquad & V = \{X, Y, Z\}, \quad F = \{F_X, F_Y, F_Z\} \\
& X := F_X(U_X) \\
& Y := F_Y(X, Z, U_Y) \\
& Z := F_Z(X, U_Z) \qquad (1.11)
\end{aligned}
$$

This particular DAG corresponds to a well known example concerning what was initially incorrectly interpreted as gender bias against women in student admissions at Berkely University (Bickel et al., 1975). That paper shows that while significantly more men were admitted to the university, when department types were taken into account, 6/85 departments were biased against men and 4/85 biased against women and no bias was discernible in any of the other departments.

In this model (1.11) X is *gender*, Z the *department choice* and Y the *admission outcome*. Since the functions give a hint of which variables are causally related, causation can be drawn in a graph by drawing arrows from the variables on the right side of each function to the variable on the left side. When doing so in the example, the corresponding graph is the same as depicted in Figure 1.1. Note that only the observed variables are depicted, and the U variables are omitted for simplicity with the assumption that they are independent. Graphical models aid the understanding of causality and they ease the use of joint probability distributions. As in classical statistics, probability distributions are important for causality to overcome uncertainty of the exact answer, by instead considering the probabilities of each possible state a variable can have. Following the graphical representation, when two variables in a model are conditionally independent, i.e. $X \perp\!\!\!\perp Y$ the following equations on the probabilities hold:

$$P(X) = P(X|Y) \; and \; P(Y) = P(Y|X) \tag{1.12}$$

The joint distribution of all variables in a model can be calculated by the product of all conditional distributions of families in the graph. This is expressed in the following rule of product decomposition (Pearl et al., 2016):

$$P(x_1, x_2, x_3, \dots, x_n) = \prod_i P(x_i|pa_i) \tag{1.13}$$

where i runs over all nodes and pa of i represents the parents of the node x_i. With the student admission problem, the probability of a female student being admitted to the engineering department, can be described by simply multiplying P(Female), P(Engineering department|Female) and P(Admission outcome|Female, Engineering department). More generally:

$$P(X = x, Y = y, Z = z) = P(Y = y|X = x, Z = z)P(Z = z|X = x)P(X = x) \tag{1.14}$$

In the Berkeley student admission problem, *gender* was observed as a causal effect of admission outcome. However, when the choice of department was added with an incoming edge from *gender* and outgoing edge to *admission outcome* in the model, which is called a *chain* structure, the effect disappeared to a great extent. In the causal graph, *Department Choice* is referred to as *mediator* between *Gender* and *Admission outcome*.

Beside *chains*, two more interesting structures in a causal model are *forks* and *colliders*, see Figure 1.2. A *fork* structure consists of a node that is a common cause of two other variables, the parent node often called a *confounder*, whereas a collider node is a common effect of two variables, the child node being the *collider*.

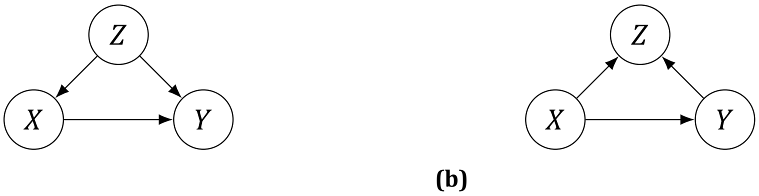


Figure 1.2 Two graph structures. (A) The middle node (Z) being a fork/confounder or in (B) a collider.

1.2.1 Model Discovery and Validity

To validate the SCM is difficult. In practice causal hypotheses, i.e. DAG topologies and functional dependency, are set up in such a way as to enable *falsification*, if sufficient appropriate data are available; most commonly the model can merely *not be falsified* which is evidently **not** the same as validation or verification. For model discovery and validity, dependencies between variables are of great importance. Following Pearl et al. (2016), the following rules on dependencies hold for the different structures:

- Chain structures (Figure 1.1): Z and X are dependent, Y and Z are dependent and X and Y are dependent.
- Fork structures (Figure 1.2a): Z and X are dependent, Z and Y are dependent and X and Y are dependent.

- Collider structures (Figure 1.2b): X and Z are dependent, Y and Z are dependent and X and Y are dependent given Z.

Thus, while it is possible to condition on mediators and forks without changing the dependencies between X and Y, conditioning on a collider Z opens up a path between X and Y and they become dependent. In more complex graph structures the notion of d-separation, where d denotes ”directional”, can be used to discover conditional independencies (Pearl, 2009):

Definition 1. ***d-separation:*** *Two variables X and Y are d-separated by a set of nodes S if all paths between X and Y are blocked by S. Paths are blocked when either:*

1. *Set S is in a fork or chain between X and Y*
   $X \ldots \leftarrow S \rightarrow \ldots Y$ *or* $X \ldots \rightarrow S \rightarrow \ldots Y$ *or* $X \ldots \leftarrow S \leftarrow \ldots Y$
2. *There is a collider (O) present between X and Y that is not part of set S, nor are any of its descendants*
   $X \ldots \rightarrow O \leftarrow \ldots Y$

Furthermore, the global Markov property is assumed, in order to relate the graph structure to probability distributions Pr in the data:

Definition 2. ***Global Markov property:*** *Pr is Markov with respect to graph G for all sets S if:*
X and Y are d-separated by S in G $\Rightarrow X \perp\!\!\!\perp Y \mid S$

with Pr being the probability distribution over the variables in the graph and with S being a set that can contain one or more variables, or be an empty set $\emptyset$. Put in words, if two variables X and Y are d-separated, X and Y are independent. On the contrary, when there is at least one path between X and Y that is not blocked, X and Y are d-connected and most probably dependent, i.e. $X \not\perp\!\!\!\perp Y$. From the Markov condition it can be assumed that the other direction holds as well. In particular, the Faithfulness assumption is agreed upon:

Definition 3. ***Faithfulness assumption:*** *P is faithful with respect to graph G for all sets S if:*
X and Y are d-separated by S in G $\Leftarrow X \perp\!\!\!\perp Y \mid S$

Lastly, the Reichenbach’s common cause principle is agreed upon (Hausman and Woodward, 1999; Peters et al., 2017):

Definition 4. ***Reichenbach’s common cause principle:*** *Assume that $X \not\perp\!\!\!\perp Y$, then either:*

1. *X causes Y*
2. *Y causes X*
3. *There is a hidden variable, also called ”confounder” as common cause (fork)*
4. *A combination of the above*

This simply confirms the idea that when two variables are dependent, there must be a causal reason for this. Using background knowledge and independence testing on the data, causalities can be discovered in the graph on the basis of the Markov and Faithfulness assumptions (Pearl, 2009). For example, independence tests on a dataset with $V = \{X, Y, Z\}$ might reveal: $X \perp\!\!\!\perp Y | Z$. Thus, X and Y must be d-separated by set Z, so the possible graphs using Definition 1 are: $X \leftarrow Z \leftarrow Y$, $X \rightarrow Z \rightarrow Y$ and $X \leftarrow Z \rightarrow Y$. Other algorithms have been constructed as a fast way to find independencies within the data (Spirtes et al., 1993).

Other techniques used for causal discovery include *regression methods* and *invariant causal prediction* (Peters et al., 2017). With two random variables X and Y, for the regression method, first X is regressed on Y and the independence between the residuals and X is checked. Then Y is regressed on X and check independence between the residuals and Y is checked, to see whether $X \rightarrow Y$ or $X \leftarrow Y$ is more suitable given the data. The overall maximum likelihood of different models can also be calculated by regressing all nodes on their parents and calculating the variance in residuals. Invariant causal prediction on the other hand, splits the data into different environments and infers causal relationships from which parent → child relationships stay invariant over the course of different environments (Peters et al., 2017).

1.2.2 Interventions

An SCM can be seen as a set of instructions to generate a distribution from noise variables. Using the SCM, the data can be manipulated and the new distribution studied. By doing manipulations, i.e. interventions, it becomes possible to answer questions that go beyond studying observations only and to study the effect of hypothetical actions. For example, by putting one of the variables to a fixed value, also called the do-operator. Under the do-operation $X := x$, a new model M' is constructed:

$$M' = M[X := x] \tag{1.15}$$

The corresponding causal effect can be described as the effect of the do-operator on the outcome variable Y, also called the average treatment effect f_{treat}:

$$f_{treat} \equiv E_{M[X:=1]}[Y] - E_{M[X:=0]}[Y] \tag{1.16}$$

where $E[Y]$ denotes the expectation of the outcome Y.

The do-operator performs a surgery on the model: by fixing the value of a variable, the incoming edges from the parents are deleted. Conditioning on a variable on the other hand, only changes the way the data are examined, by focusing on only the cases of a specific group. If there is a difference in probabilities after the do-operator and after conditioning, as in most cases, this is called confounding:

$$P\{Y = y|do(X := x)\} \neq P\{Y = y|X = x\} \tag{1.17}$$

Thus, these terms are not interchangeable. However, the effect the do-operator has in terms of conditional probabilities can be estimated by conditioning on the parent nodes (Pa) of node X, also called the adjustment formula:

$$P\{Y = y|do(X := x)\} = \sum_{z} P\{Y = y|X = x, Pa = z\}P\{Pa = z\} \tag{1.18}$$

However, one must be careful to not condition on a collider, which will create a dependency between the adjacent nodes, as discussed above. To easily find deconfounding variables that can safely be controlled for, the *backdoor criterion* is used (Pearl, 2009): they must not have a backdoor path present between the nodes. A practical solution for confounding variables is to use a randomized controlled trial. This way, all incoming edges are deleted, which naturally closes all backdoor paths and potentially confounded paths. This is not always possible, and also not always necessary.

In the particular example of discrimination, relevant for subsequent sections of this paper, indirect paths with a mediator can contribute to discrimination by unobserved variables. For instance in the case of the Berkeley admission problem, advertisements of the engineering department that are focused more on men than women, or fear of woman applicants to get rejected in the engineering department, might be unobserved confounding variables (Barocas et al., 2019). Thus, before conditioning on a set of variables, it is important to envision all indirect paths that might hold discrimination. The often-complex social structure behind discrimination requires precise path inspection. Furthermore, these different paths can either be considered socially acceptable or unacceptable/discriminating factors. Estimating the indirect effect by disabling the direct effect of a sensitive attribute on the outcome is not possible, but counterfactuals provide a way to estimate these path specific effects.

1.2.3 Counterfactuals

Counterfactuals are of a higher level than observations and interventions. They can answer questions concerning situations that cannot be observed in the real world, hence the name counter-factual. Consider the following example: you have a headache and decide to take a pill. After some time the headache disappears. The counterfactual question would be: would my headache have disappeared if I were not to have taken a pill? As there can only be one realisation - either you take the pill or not - it is not possible to know the answer to this situation. Structural causal models give a way to answer such questions. Generally, the observational evidence that the headache disappeared, holds some information and modifies the most appropriate distribution function for the stochastic variation of the variables in the model. This consequence is often referred to with the short-hand formulation “biasing of the noise”. So a first step is to condition on the event and calculate the new biased distribution of noise variables U'. Secondly, the do-operation $X := x$ is performed, as the treatment variable needs to be fixed on not taking a pill. This results in a new structural causal model M'. Thirdly, target counterfactual is computed $Y_{X:=x}(E)$ with the new U' as initial seed in M'. The definition from Barocas et al. (2019) will be used:

Definition 5. ***Counterfactual in a structural causal model:*** *Given a structural causal model M, an observed event E, an action X:=x and a target variable Y, the counterfactual $Y_{X:=x}(E)$ is defined by the following three step procedure:*

1. ***Abduction****: Adjust noise variables to be consistent with the observed event. Formally, condition the joint distribution of $U = (U_1, \dots, U_d)$ on the event E. This results in a biased distribution U'.*
2. ***Action****: Perform do-intervention $X := x$ in the structural causal model M resulting in the model $M' = M[X := x]$.*
3. ***Prediction****: Compute target counterfactual $Y_{X:=x}(E)$ by using U' as the random seed in M'.*

Counterfactuals can quantify path specific effects. Confounded paths can be eliminated by using the do-operator $E[Y|do(X := X)]$. For mediating paths, as in Figure 1.1 counterfactuals are used to decompose the causal effect into the direct effect of $X \rightarrow Y$ and indirect effect of $X \rightarrow Z \rightarrow Y$. The direct effect is quantified by setting the mediator to a specific level:

$$E[Y|do(X := 1, Z := z)] - E[Y|do(X := 0, Z := z)] \tag{1.19}$$

This is also called the controlled direct effect. Besides, the direct effect can be calculated by using the natural direct effect:

$$E[Y_{X:=1,Z:=Z_{X:=0}} - Y_{X:=0,Z:=Z_{X:=0}}] \tag{1.20}$$

Hereby, the mediator Z can still vary as if no treatment occurred: $X := 0$.

The natural direct effect generates a notion of the natural indirect effect:

$$E[Y_{X:=0,Z:=Z_{X:=1}} - Y_{X:=0,Z:=Z_{X:=0}}] \tag{1.21}$$

Hereby, the treatment is fixed to no treatment and the effect is examined of treatment versus no treatment on the mediator value. With counterfactuals it becomes possible to compute path-specific effects and design decision rules that eliminate path-specific effects that are undesirable, for example based on assumptions regarding discrimination. Before considering normative fairness measures and counterfactual discrimination practices in further detail, first a useful technique to do causal inference is discussed.

1.3 Causal Effect Variational Autoencoder

Exact inference can be difficult when a causal model consists of a lot of variables, including possibly some unobserved ones. Therefore, approximate inference is proposed as a method for inferring the causal relationships and unobserved variables. In particular, the Causal Effect Variational Autoencoder (CEVAE) method can be applied (Louizos et al., 2017). A Variational Autoencoder is a latent variable model that learns to represent the input data via an encoder and a decoder (Kingma and Welling, 2013; Rezende et al., 2014). When the input of the model is of a certain dimension, the hidden layer, also called a bottle neck, must be of a smaller dimension and the output layer is again of the same dimension as the input layer. Both the encoder and decoder are neural networks. The encoder encodes the input layer to the latent space Z, which is a distribution over Z. The decoder network reconstructs the input from latent space Z to the output layer. The network learns by maximizing the variational lower bound of the likelihood of observing the data in the current model. As the network learns to represent the data, the latent space holds valuable information in a lower dimensional space, i.e. high-level features, that can still accurately represent the input data. Many studies have demonstrated the validity of the method in the past (Chung et al., 2015; Gregor et al., 2015; Louizos et al., 2015).

In the CEVAE in particular, the latent space represents unobserved variables, and the decoder represents the data generation process by a causal graph (Helwegen et al., 2020; Louizos et al., 2017). Hence, the decoder is based on the structure of the causal graph, whereby the nodes/variables are represented as probability distributions. The edges are represented as neural networks in order to capture the non-linear relationships between the variables. Following Helwegen et al. (2020); Louizos et al. (2017), stochastic optimization of the network by maximizing the variational lower bound is done via the following equation:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{reg} \tag{1.22}$$

so that $\mathcal{L}$ is the combined loss of the reconstruction and regularization loss terms. The reconstruction loss considers the log-likelihood of the probability distributions given the observed variables and is calculated by:

$$\mathcal{L}_{rec} = \sum_{i=1}^{N} \mathbb{E}_{q(z_i|x_i,y_i)}[\log\ p(x_i|pa_i) + \log\ p(y_i|x_i,z_i)] \tag{1.23}$$

where i refers to an instance, z_i to the latent space z, y_i to the outcome of the current instance, x_i to the other variables and pa_i the parents of these variables in the model. The regularization term controls the variation between the true posterior and the approximate posterior. The regularization and loss is calculated as in eq. 1.24.

$$\mathcal{L}_{reg} = \sum_{i=1}^{N} \mathbb{E}_{q(z_i|x_i,y_i)}[\log\ p(z_i) - \log\ q(z_i|x_i,y_i)] \tag{1.24}$$

Using reparametrization, the optimization of the network can be done using a stochastic gradient descent method such as e.g. Adam (Kingma and Ba, 2014).

2 Application: counterfactual fairness

In previous studies, different fairness metrics on (algorithmic) decisions have been proposed and tested. Based on the review of Verma and Rubin (2018), an overview will be given. For simplicity, here the focus is on binary classification. The usual statistical measures are employed to quantify the performance of an algorithm that produces a predictied probability score S for an instance to belong to a class: True Positive (TP) is an instance that is accurately classified as positive, False Positive (FP) is an instance that is wrongly classified as positive, False Negative (FN) is an instance that is wrongly classified as negative and True Negative (TN) is an instance that is accurately classified as negative.

The following proportion measures are considered: Positive Predictive Value (PVV), True Positive Rate (TPR), False Discovery Rate (FDR), False Positive Rate (FPR), False Omission Rate (FOR), False Negative Rate (FNR), Negative Predictive Value (NPV) and True Negative Rate (TNR), see Table 2.1 for their corresponding formulas (Verma and Rubin, 2018).

Table 2.1 A confusion matrix containing statistical measures on the performance of an algorithm.

	Actual - Positive	Actual - Negative

Predicted - Positive	True Positive (TP)
	PPV = $\frac{TP}{TP+FP}$
	TPR = $\frac{TP}{TP+FN}$
		False Positive (FP)
		FDR = $\frac{FP}{TP+FP}$
		FPR = $\frac{FP}{FP+TN}$

Predicted - Negative	False Negative (FN)
	FOR = $\frac{FN}{TN+FN}$
	FNR = $\frac{FN}{TP+FN}$
		True Negative (TN)
		NPV = $\frac{TN}{TN+FN}$
		TNR = $\frac{TN}{FP+TN}$

The fairness conditions, based on these measures, that can be distinguished are listed below (Verma and Rubin, 2018). The first nine are based solely on the aforementioned statistical measures. Fairness measures 10–13 consider a probability prediction, 14–16 consider similarities between instances/subjects and 17–20 elaborate fairness with respect to a causal graph. Groups are formed on the basis of a sensitive attribute. For example, the performance of the algorithm on people with a Dutch ethnicity can be compared to the performance on people with a non-Dutch ethnicity.

1. **Demographic parity/Statistical parity**: $\frac{TP+FP}{total}$ should be equal between groups.
2. **Conditional demographic parity**: same as Demographic parity, but with some legitimate attributes that are allowed to influence the prediction and should be controlled for.
3. **Predictive parity/Precision parity**: PPV should be equal between groups, and consequently FDR should.
4. **Predictive equality/False positive parity**: FPR should be equal between groups, and consequently TNR should.
5. **Equal opportunity/True positive parity**: FNR should be equal between groups, and consequently TPR should.
6. **Equalized odds/Positive rate parity**: TPR and FPR should be equal between groups (combination of the previous two).
7. **Conditional use accuracy equality**: PPV and NPV should be equal between groups.
8. **(Overall) accuracy parity**: $\frac{TP+FN}{total}$ should be equal between groups.
9. **Treatment equality**: $\frac{FP}{FN}$ should be equal between groups.
10. **Calibration**: TP probability should be equal for all probability scores S.
11. **Well-calibration**: same as Calibration, plus the TP probability should match the probability score S for all groups.
12. **Balance for positive class**: average predicted probability score S should be equal for subjects in the positive class.
13. **Balance for negative class**: average predicted probability score S should be equal for subjects in the negative class.
14. **Causal discrimination**: any two subjects with the exact same attributes X, but a different group, should have similar classification.
15. **Fairness through unawareness**: no sensitive attribute should be explicitly used in the decision-making process.
16. **Fairness through awareness**: similar individuals (via a distance metric) should have similar classification.
17. **Counterfactual fairness**: given a causal graph, the predicted outcome in the graph should not depend on a sensitive attribute, see Definition 6.
18. **No unresolved discrimination**: given a causal graph, there should be no path from the sensitive attribute to the predicted outcome, except via resolving variables, the variables deemed non-discriminatory.
19. **No proxy discrimination**: given a causal graph, there should be no path from the sensitive attribute to the predicted outcome that is blocked by a proxy variable.
20. **Fair inference**: given a causal graph, and paths that are classified as legitimate or illegitimate, there should be no illegitimate paths from the sensitive attribute to the predicted outcome.

Some fairness conditions are mutually exclusive so it is necessary to make a choice which one to optimise in any given application. For example, consider the USA discrimination laws *disparate treatment* and *disparate impact*. Disparate treatment states that all people with equal ability should have the same opportunity, i.e. equality of opportunity. Disparate impact states that inequality in the outcome should be minimized. But minimizing inequality in the outcome might result in a different treatment for similar people. This collision of preferences encoded in law has produced serious problems with the legal mechanism intended to minimise discrimination in the USA in the past. A second issue is that perfect classification does not exist. Choosing which is the right fairness measure depends on which mistakes are most tolerated for the given problem. When detecting fraud, False Positives are highly disapproved, much more so than for instance with a classifier that scans CVs for job hiring. One example in the domain of education, is the Texas 10% rule, or Texas House Bill 588, that guarantees admission to state-funded universities for the best 10% of each class, resulting in more diverse university classrooms (U.S.C., 1996). As for the Netherlands, out of the four main functions of education one is to stimulate equal opportunities (Werfhorst, 2007). For example, more money is distributed to schools in deprived areas (Min. OCW, 2017). For more information on the Dutch laws relevant to this problem, some excerpts are given in the Appendix I Dutch Laws on Equality and Education.

As multiple studies have pointed out, non-causal fairness measures do not take into account all variables and examine the fairness of their relationships. Consequently, for this study we focus on causal fairness and specifically, the individual level counterfactual fairness is proposed as a valid metric. However, as mentioned earlier, this is time and context dependent and current shared social attitudes. as well as legal and governance context, should help with an ultimate choice before application.

2.1 Defining counterfactual fairness

Counterfactual demographic parity or counterfactual fairness (Kusner et al., 2018), is based on its observational counterpart called conditional demographic parity. Counterfactual demographic parity states that the distribution of the predictor should stay the same in a different world where the sensitive attribute is different, i.e.: $\hat{Y}_{A:=a}(E)$ and $\hat{Y}_{A:=a'}(E)$ should follow the same distribution. On the individual level, when the value of the sensitive attribute of a person is changed, this should have no effect on the outcome of the model. The following notion of counterfactual fairness is used (Kusner et al., 2018):

Definition 6. ***Counterfactual Fairness.*** *Predictor $\hat{Y}$ is counterfactually fair if for individuals with observed variables X=x and sensitive attribute A=a the following is satisfied:*

$$P(\hat{Y}_{do(A=a)} = y|X = x, A = a) = P(\hat{Y}_{do(A=a')} = y|X = x, A = a)$$

for all y and any value of a', where $P(\hat{Y})$ is the probability distribution of the predictor and $do(A = a')$ is the intervention of changing the sensitive attribute to a'.

For example, if the outcome variable Y refers to salary categories and sex is the sensitive attribute then changing in the available dataset the value of the variable sex from ‘male’ to ‘female’, or vice versa, while keeping other variables invariant should yield the same probability of model outcome. A violation of this would indicate that reality appears not to conform to the counterfactual fairness criteria.
Importantly, to construct valid counterfactuals, flipping the sensitive attribute should also adjust downstream features. Only changing the value of the sensitive attribute and letting the other features remain the same, would wrongly assume that there are no descendants of the sensitive attribute in the graph. This is therefore distinct from adjusting only marginal propensities in a model. This notion of fairness is satisfied by default when only non-descendants of sensitive attributes in the causal graph are used to predict the outcome.

One important thing to mention is the difficulty of validating the structural causal model that includes a sensitive attribute like ethnicity. Ethnicity can either be seen as a biological or a social construct. As social construct, it is difficult to find a set of nodes to represent the influence of society on ethnicity. Besides, the social constructs lack modularity, as structural causal models are systems of reduced complexity. Furthermore, ethnicity as a feature might be unstable, as the world changes over time. As discussed earlier, this effect cannot be taken into account in an acyclic graph. As a final remark, people that belong to a specific group might change their behavior based on their agreement with the group they are assigned to. This is also called a looping effect, which changes the model over time (Hacking, 2006).

2.2 (Meta-)data used

2.2.1 Qualitative

The causal model in this study was constructed to be as closely related to the real decision problem as possible, concerning school level selection for secondary education. In order to reduce the researcher bias, being in the present case the potential bias of ourselves as authors, concerning knowledge on the subject, studies on bias in the Dutch selection procedure were examined. A systematic literature search was performed using the search engines Google Scholar and PubMed. Search terms included the following words: ”bias”, ”inequality”, ”selection procedure”, ”tracking”, ”school system”, ”Dutch”, ”the Netherlands”, ”ethnicity”, ”gender” and ”social economic status”. For selecting the articles, there was a focus on studies that analyzed inequalities in the selection procedure. Besides, only studies on the Dutch school system were considered, or studies in countries with a similar education structure. This resulted in a total selection of 18 studies.

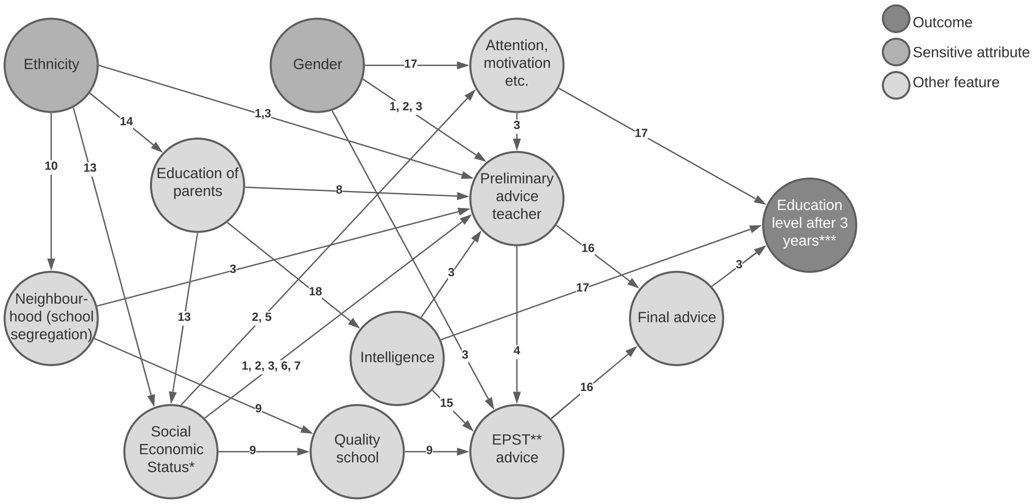


Figure 2.1 A theoretical model on bias in the Dutch education system. *Can be based on income of parents, education of parents, estimated value of the house and the occupation of parents. **EPST = End-of-Primary-School-Test. ***Track mobility and education level are again influenced by sensitive attributes [11], [12]. References: 1. Timmermans et al. (2018), 2. Geven et al. (2018), 3. Lek (2020), 4. Gentrup et al. (2020), 5. Boone and Van Houtte (2013), 6. Ditton et al. (2005), 7. Dumont et al. (2019), 8. Timmermans et al. (2015), 9. Agirdag et al. (2013), 10. Boterman (2019), 11. Tieben (2011), 12. Tolsma et al. (2007), 13. CBS (2018), 14. CBS (2021), 15. Hop and van Boxtel (2013), 16. Ministerie van Onderwijs Cultuur en Wetenschap (2021), 17. Spinath et al. (2014), 18. Marioni et al. (2014).

With this a theoretical model of the Dutch education system is constructed, see Figure 2.1. The nodes in the figure illustrate the variables that play a role, consisting of the sensitive attributes (ethnicity and gender), the outcome of the model (the ultimate education level, three years after the selection procedure) and other identified features. The arrows in the graph represent some sort of a *causal link* (in the sense of section 1.2) between the predecessor and the successive node. The numbers in Figure 2.1 are references to the studies that indicated these effects.

The Dutch secondary school system is differentiated in a number of types, although there are provisions for pupils to switch between these types (see also Fig. 2.6 below). The distinct types have the abbreviations 'vwo' (a pre-university track), 'havo' (a pre-further education track), and 'vmbo' (a pre-vocational track). Within the 'vmbo' track there is still a distinction between a somewhat more theory oriented ('vmbo-t') or somewhat less theory oriented ('vmbo-k') sub-track or an intermediate sub-track ('vmbo-g') and finally also a more basic skills oriented sub-track ('vmbo-b'). Often in secondary schools, the different tracks, or a subset of these, are offered co-located in the same building. In a minority of secondary schools there are even *mixed classes* for the lower years, so that the differentiation between two tracks, for instance the vwo and havo track, is postponed compared to what is generally done in the Netherlands.

The literature analysis identified numerous potential causes for the inequalities within the Dutch education system, with varying levels of supporting statistical evidence. To start with, teachers might be biased, either in an implicit or explicit way (Geven et al., 2018; Timmermans et al., 2018). Some teachers might have implicit prejudices against certain social or ethnic groups, or they might take into account a sensitive attribute itself or a proxy variable for constructing the advice, such as the education of parents or situation at home (Ditton et al., 2005). This effect might have increased after the change in the selection procedure, whereby the teacher's estimate gained a higher priority over the EPST test result. Besides, it is observed that parents that attained higher education have a more powerful position in conversations with the teacher, and object more often (Dumont et al., 2019; Timmermans et al., 2018).

Parents that attained higher education also seem to be more involved in school activities. Some studies, but not all, revealed a higher involvement in school practices (Geven et al., 2018). Besides, higher educated parents have a higher admiration and have higher expectations for their children (Geven et al., 2018). Parents also impact the decision of primary school, which appears in the high level of school segregation, especially in the cities of the Netherlands (Vogels et al., 2021). The differences in quality of the school and the classroom composition might again affect the inequalities in education, albeit the effects of classroom composition can be both positive and negative (Geven et al., 2018).

Moreover, the individualistic culture and the idea that each child should be maximally stimulated in order to optimize growth, leads to more differentiation within the classroom and between schools (Denessen, 2017; Van de Werfhorst, 2011). Secondary schools use this differentiation as advertisement to attract parents and their children. This is one of the reasons why, despite the advice of the Dutch educational inspection to counteract inequalities, numbers of broad schools have been decreasing over the years (Vogels et al., 2021).

The Dutch education system is known for its early differentiation at the age of 11 or 12. The early selection has its roots in the past century (the so-called Mammoetwet in the 1960s) and common social attitudes around that time. but now appears outdated. At such a young age, the pupils are mainly a reflection of their parents (Shavit and Blossfeld, 1993). In comparison with countries like Austria, Germany and Switzerland, the Netherlands have a highly differentiated secondary school system (Van de Werfhorst, 2011). From comparative studies, we know that a high level of differentiation is associated with more inequalities between social and ethnic groups (Van de Werfhorst and Mijs, 2010), and less transfer ability to higher education. It is debatable whether the high level of differentiation stimulates the learning environment in the class at all (Van de

Werfhorst, 2011). It seems that at least the gain in knowledge of children in higher levels do not outweigh the decline in learning of children in lower levels. Reasons might entail the waste of talents and learning loss in lower tracks. High differentiation is also associated with a lower desire to continue higher education and lower societal participation (Van de Werfhorst, 2011).

Another cause is the difference in educational and monetary resources that children from different social and ethnic groups have (Aalders et al., 2020; Vogels et al., 2021). Parents who attained higher education are more capable of supporting and helping their children with learning. This became especially clear during the COVID pandemic (Bol, 2020). Besides, wealthier families can afford educational trips during free time and provide their children with extra training (Kuyper and Van der Werf, 2012). Even EPST trainings to get a higher advice exist, although this is against the advice of the EPST test makers (Bisschop et al., 2019; PO-raad, 2019).

Yet another cause is the language level of the pupils. Pupils that have a family with a migration background may have more difficulty in keeping up with the Dutch language. This again affects the involvement of parents and their capability to support their children with learning. For pupils that speak a different language at home for the first four years of their lives, the selection at the age of 11 or 12 might be too early to reach the appropriate language level. The current focus of education on the Dutch language makes it a difficult problem. Other research revealed that performance expectations are lower for pupils with a migration background or low social economic status (Dijks et al., 2020; Geven et al., 2018). Low expectations are negatively correlated with growth, by self-fulfilling prophecy and fixed mindset. This is also the case for relatively low advices, as the children know the advice of the school before making the test, and believe the teacher knows best (Timmermans et al., 2013).

It is important at this point to emphasise that the analysis in the paper is exploratory. The primary purpose here is to demonstrate the technique and therefore a choice is made to work with a pared-down version of the full causal diagram. A more complete analysis would certainly appear to be worthwhile but is beyond the scope of this research. The actual (un)fairness of the Dutch education system should only be judged on the basis of a more complete causal model. From all the different factors that play a role, it can be concluded that the problem is complex. It is important to identify all factors and determine the size of their impact. For an overview of all the identified factors, see Figure 2.1. In what follows, the most important factors are summarized in the causal graph of Figure 2.4, which is in essence a pared-down version of the graph presented in Figure 2.1. Since ethnicity seemed to have a larger impact on the selection procedure compared to gender, ethnicity is picked as a sensitive attribute. The causal graph includes the following variables and their causal relationships: sensitive attribute A, social economic status S, language level L, preliminary advice teacher T, EPST advice E, the unknown variable that holds intelligence U and the outcome variable Y, which is the educational level 3 years after the selection procedure. Most factors cannot be influenced by the students themselves. It is important to discuss with teachers and experts which influences are deemed fair and which are deemed unfair. For this study, multiple different combinations of pathways were proposed for a fair prediction.

2.2.2 Quantitative

For this study, Dutch pupils that attended the selection procedure in school years 2015/2016, 2016/2017 and 2017/2018 were included (n = 297.035; mean age at the start of the selection procedure = 11.96 $\pm$ 0.45). Pupils were only included if three years later, they were enrolled in school year three of Secondary Education and they were not in a *mixed class*. Both information on the pupils during the selection procedure and their educational level after three years was extracted from the CBS databases. In particular, the following attributes were gathered: the *sensitive attribute*, *social economic status*, *language level*, the *teacher's preliminary advice*, the *EPST advice*, the *final advice* and the *education level after 3 years*. Pupils that had missing data in one of the target attributes were dropped[1], except for students with only one missing datapoint for either education of the mother or education of the father. More details on the attributes are provided below.

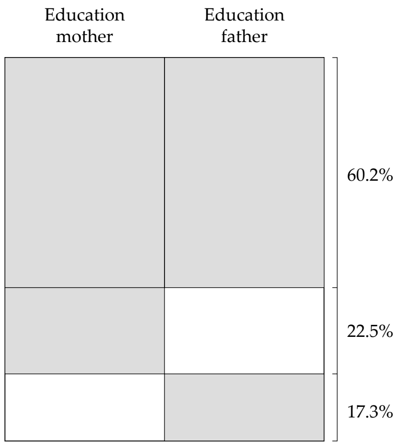


Figure 2.2 Missing data imputation. Size and distribution of the missing data: grey boxes represent data that are present and white boxes data that are missing.

- *Sensitive attribute*:
  - **Ethnicity**: for second-generation immigrants, i.e. at least one parent is not born in the Netherlands, the country of birth of the mother was selected, unless the mother was born in the Netherlands, otherwise the country of the father was selected. For first-generation immigrants, the country of birth was selected. The resulting countries of origin were first divided into the four largest ethnicity groups within the Netherlands: Dutch (1), Moroccan (2), Surinamese, Dutch Antillean or Aruban (3) and Turkish (4). The remaining countries were divided into three groups: from EU-15 countries[2] and developed economies (5), from new EU countries and economies in transition (6) and from other countries (7), resulting in a total of seven categories.
- *Social Economic Status* (SES). The following indicators for SES were included in the vector:
  - **Highest attained education of both the mother and the father**: eight categories can be distinguished: Primary Education (1), ”vmbo-b”/”vmbo-k”/”mbo 1” (2), ”vmbo-g”/”vmbo-t”/junior classes ”havo”/junior classes ”vwo” (3), ”mbo 2”/”mbo 3” (4), ”mbo-4” (5), ”havo”/”vwo” (6), ”hbo” bachelor/”wo” bachelor (7) and ”hbo” master/”wo”

[1] The study started with a sample of (close to) 531,5 thousand pupils and was reduced to (just over) 297 thousand pupils, due to the selection criteria and (multiple) missing values. This reduction is important for selection bias purposes. This is acceptable for demonstrating the technique, but a more complete analysis ought to investigate the effect of such listwise deletions on the model, and fairness measures, outcomes

[2] The EU-15 countries, as defined by CBS are: Belgium, Germany, Denmark, Finland, France, Greece, Ireland, Italy, Luxembourg, the Netherlands, Austria, Portugal, Spain, United Kingdom and Sweden.

master/doctor (8). The categories were encoded as dummy variables with the first category being dropped[3]. Only when one out of both variables for education mother and father was missing, this attribute was imputed based on all the other variables, to let the variables stay correlated. See Figure 2.2 for the size and distribution of the missing data. The imputation technique used was 'predictive mean matching', which selected five similar entries and randomly picked one of the values for imputation. This process resulted in five datasets that were all used for subsequent analysis in the experiment (Van Buuren, 2018). See Figure 2.3 for the results of the data imputation. No differences in results were found between the datasets.
  - **The estimated value of the house**: the value of the house the pupil was living during the selection procedure, under the Dutch version of the *Valuation of Immovable Property Act*, i.e. the *Wet Waardering Onroerende Zaken* (WOZ). The data were controlled for the annual increase in WOZ value. By Winsorizing the data, the influence of outliers was minimized: values higher than 1.000.000 were assigned the value 1.000.000, which affected 0.72% of the data. Afterwards, the data were normalized.
  - **Percentile disposable income**: the income of the household of the house where the pupil was living during the selection procedure. The percentile groups were normalized by default.
  - **The type of income of the mother and father**: employee (1), director and major stakeholder (2), self-employed entrepreneur (3), self-employed (4), family worker (5), unemployment benefit (6), welfare benefit (7), other social benefits (8), employment and support allowance (9), pension (10), studying with income (11), studying without income (12), without income (13). The categories were encoded as dummy variables with the first category being dropped.
- *Language level*
  - **Reading skills**: three indicators of the Dutch language level can be distinguished: lower than 1F, 1F and 2F. The categories were encoded as dummy variables with the first category being dropped.
  - **Language skills**: the same categories as reading skills were used.
- *Preliminary advice of the teacher*. The eleven types of advices are: ”vmbo-b”, ”vmbo-b/k”, ”vmbo-k”, ”vmbo-k/g”, ”vmbo-g”, vmbo-g/t”, ”vmbo-t”, ”vmbo-t/havo”, ”havo”, ”havo/vwo” and ”vwo”. The categories were encoded as dummy variables with the first category being dropped.
- *EPST advice*. The nine types of advices are: ”vmbo-b”, ”vmbo-b/k”, ”vmbo-k”, ”vmbo-g”, vmbo-g/t”, ”vmbo-t”, ”havo”, ”havo/vwo” and ”vwo”. The categories were encoded as dummy variables with the first category being dropped.
- *Final advice*. The same categories as *Preliminary advice of the teacher* were used.
- *Education level after three years*. The six types of education level are: ”vmbo-b”, ”vmbo-k”, ”vmbo-g”, ”vmbo-t”, ”havo” and ”vwo”. This ordering is also used when referring to higher or lower educational attainment.

Accordingly, each record in the dataset represented a pupil and consisted of a vector of size 85.

2.3 Data analysis

[3] The first category/column is being dropped for categorical variables to avoid multicollinearity within the data: highly correlated variables, with a single variable that can be predicted based on other variables.

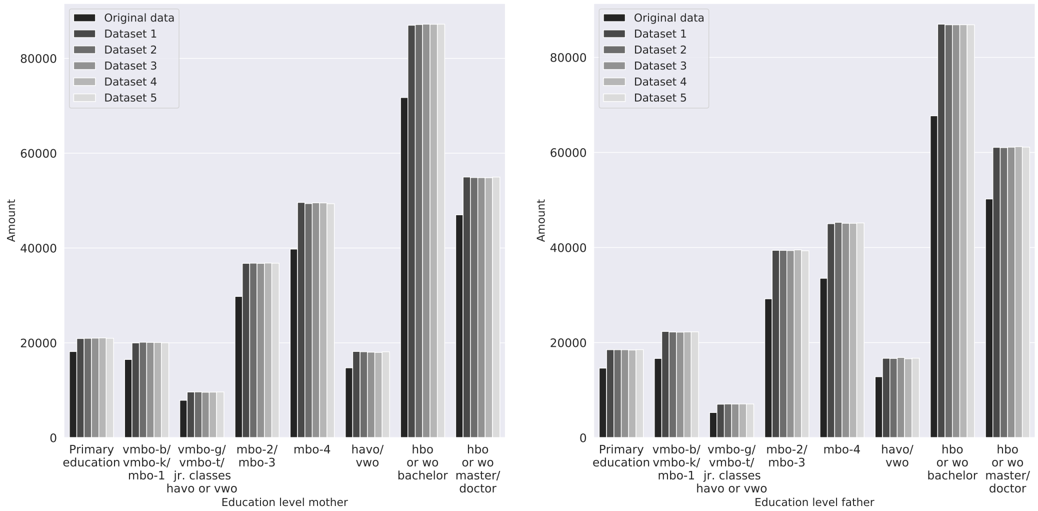


Figure 2.3 Results of missing data imputation. Left side: education mother, right side: education father. The amount per education level is depicted for the original dataset and the five imputed datasets.

2.3.1 simplifying the graph

The data analysis was performed on the causal graph depicted in Figure 2.4. The results of literature search, described in section 2.2.1, produced a more detailed graph (Fig. 2.1) which for convenience and tractability of the analysis has been simplified somewhat. The different variables are denoted as follows:

- A: sensitive attribute (ethnicity)
- L: language level
- S: social economic status
- T: teachers' preliminary advice
- E: EPST advice
- F: final advice
- U: unknown variable that holds intelligence
- Y: outcome educational level 3 years after the selection procedure

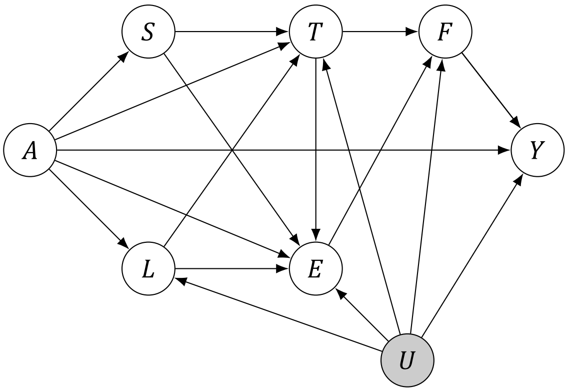


Figure 2.4 The causal graph on bias in the selection procedure, where A is the sensitive attribute (ethnicity), S is the social economic status, L is the language level, T is the teachers' preliminary advice, E is the EPST advice, F is the final advice, U is the unknown variable that holds intelligence and Y is the outcome educational level 3 years after the selection procedure. The grey node represents an unknown variable.

There is one unknown variable U, *intelligence*, hence the CEVAE is constructed to approximately infer the distribution over the latent space of the variable U and the causal relationships at the same time (Louizos et al., 2017). From now on, the latent space is denoted as latent space U. Since the true posterior depends on the variables L, T, E and F, the encoder network consists of the approximate posterior $q(u|l,t,e,f)$. The predictor Y is absent in the encoder network otherwise the network would be trained on information of the predictor itself and out-of-sample prediction would not be possible. The network consists of an input layer with the size of the total dimensions of L, T, E and F, a hidden layer with a dimension of 50 with exponential linear unit (ELU) activation (Clevert et al., 2015) (required for improved modelling of non-linear relationships between the input and outcome variables), and the output is restricted to be 5-dimensional and is considered a product distribution of standard Gaussians:

$$q(u_i|l_i,t_i,e_i,f_i) = \prod_{j=1}^{D_U} \mathcal{N}(\mu_{ij},\sigma_{ij}) \tag{2.1}$$

whereby i refers to a point in the dataset, i.e. a student, j to the dimension, D_U to the dimension of latent space U and μ_{ij} and σ_{ij} are the parameters of the standard Gaussian learned separately by the neural network. The decoder network consists of the model posterior $p(u)$ which is represented as a normal distribution with a mean of 0 and a standard deviation of 1, see Eq 2.2, and multiple reconstruction networks for the observed variables.

$$p(u_i) = \prod_{j=1}^{D_U} \mathcal{N}(u_{ij}|0,1) \tag{2.2}$$

After a sample of the inferred U is taken, the observational data of each variable in Figure 2.4 are reconstructed using six separate neural networks, one for each of the inputs, to approximate the parameters of the distributions. Each neural network of variable x_i with $x_i \in \{L,S,T,E,F,Y\}$, has as input the observational data of its parent nodes pa_i, a hidden layer with a dimension of 50 with ELU activation and the output has the same dimension as the variable x_i itself. For all decoder networks that reconstruct the observed variables, a TARnet structure is used (Shalit et al., 2017) to preserve the effect of the treatment, as in Louizos et al. (2017). This is extended to be k separate heads for the different groups of the sensitive attribute, as in Schwab et al. (2018). For example see Figure 2.5 for the network architecture for reconstructing variable L (language level) which has as input the 5-dimensional variable U, the four different types of A are represented in the k head structure, the network has a hidden layer of 50 nodes and each head has as output four nodes, similar to the size of the variable itself.

All binary variables in the model are parametrized as Bernoulli distributions, whereby the neural network outputs the corresponding probability using a Sigmoid function to get an outcome within the range [0,1] (Han and Moraga, 1995). Continuous variables are parameterized as Normal distributions, whereby the neural network separately outputs the mean μ and standard deviation σ. The standard deviation is forced to be positive by using the Softplus activation function (Glorot et al., 2011). The output network of the predictor Y has a Softmax activation function to get values within the range [0,1] for each possible outcome/category, i.e. the six different education levels, that sum up to one (Goodfellow et al., 2016). To calculate the reconstruction loss of Y, the highest value is selected and compared with the actual Y of the

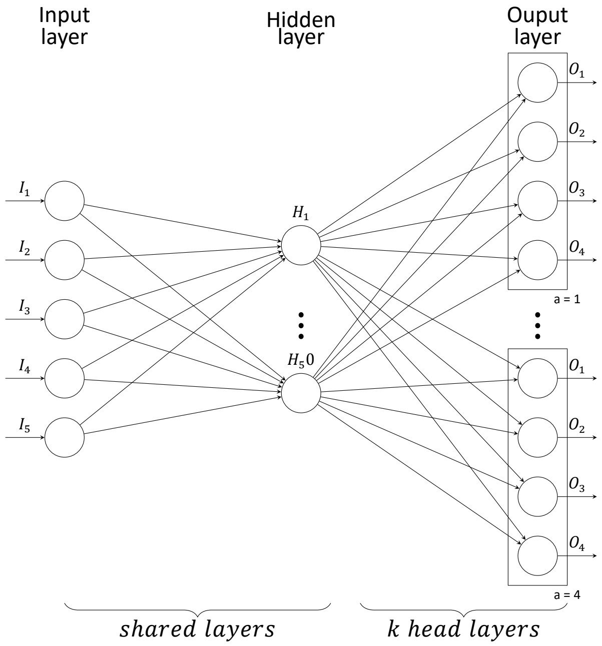


Figure 2.5 Neural network architecture for reconstructing the variable L (language level). The network has as input the 5-dimensional variable U. Four heads represent the variable A. Each head has four nodes, similar to the size of the variable itself.

batch data. Given the equations Eq. 1.22, Eq. 1.23 and Eq. 1.24, the objective of this CEVAE model becomes:

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^{N} \mathbb{E}_{q(u_i|l_i,t_i,e_i,f_i)} [&\log\ p(l_i|a_i,u_i) + \log\ p(s_i|a_i) + \log\ p(t_i|a_i,l_i,s_i,u_i) \\ &+ \log\ p(e_i|a_i,l_i,s_i,t_i,u_i) + \log\ p(f_i|a_i,t_i,e_i,u_i) \\ &+ \log\ p(y_i|a_i,f_i,u_i) + log\ p(u_i) - \log\ q(u_i|l_i,t_i,e_i,f_i)]\end{aligned} \tag{2.3}$$

The optimizer Adam (Kingma and Ba, 2014) is used with a learning rate of 0.0005 and a batch size of 500. A total of 150.000 training iterations are performed. The data are split into training and test data (10%). The model is tested on the test set every 100 times to check for overfitting.

After construction of the CEVAE model, ML models with different input selections are built, to propose different kinds of fair ML models. The different input structures are as follows: $\{U\}$, $\{U, F\}$, $\{U, L, S, T, E, F\}$ and $\{U, A, L, S, T, E, F\}$. Each model is based on part of the selected attributes and the models predicted the educational level of pupils 3 years after the selection procedure. Just as done above, the models' architecture has a hidden layer of 50 nodes with ELU activation and the output of the network has six dimensions, one for each educational level, again with Softmax activation to get a probability for each category that adds up to one. The networks were trained for 1.000 iterations with a batch size of 500. Each training loop, U was sampled from the learned CEVAE model given the batch data. Multiple models with different inputs are constructed to compare accuracy and fairness of the method. The models were optimized using RMSprop (Tieleman and Hinton, 2014) with a learning rate of 0.0005. The accuracy of each model is calculated by selecting the highest two probabilities, as is done when two subsequent educational levels are predicted in real life. If the actual outcome Y is part of them, the prediction is considered correct. Accuracy was also evaluated in terms of how secure the model is in the case of a correct and an incorrect prediction. Besides, a confusion matrix was constructed to evaluate the type of mistakes. A total of 10 repetitions are performed to measure the stability of the method.

2.3.2 preprocessing

The data are first inspected to get an indication of the quality and to get the degree of inequality within the current sample. From the 297 thousand pupils, a total of 46.2% of the pupils switched to another level within 3 years of secondary education, of which 43.4% switched one or multiple levels up. When mixed classes are not taken into account, which lead to a switch by default, 27.0% of the students switched to another education level, of which 40.9% up. See Figure 2.6 for a visualization of the switching behavior, with on the left side the proportion of pupils for the different final school advices and on the right side the education levels 3 years later.

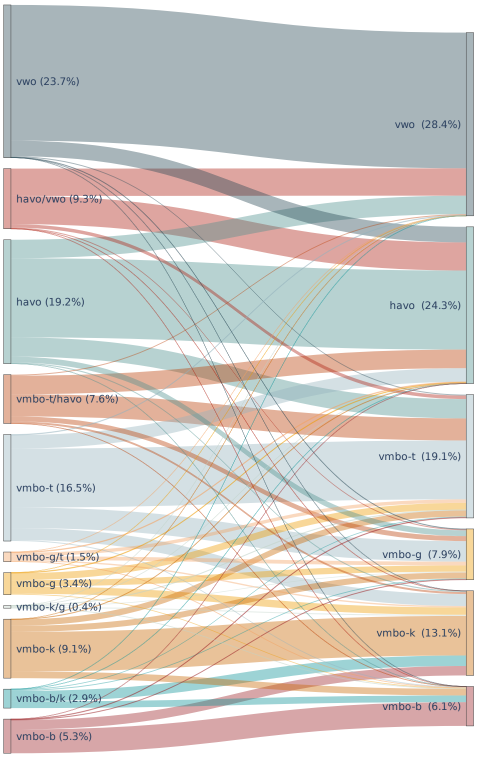


Figure 2.6 The switching behavior of students up- and downstream. Final school advices are depicted on the left. The education level students reached after 3 years are depicted on the right. The thickness of lines represent the amount of pupils.

To get an indication of the inequality within the current sample, the final school advices and education level reached after 3 years were grouped by ethnicity type. The percentage students for each of the different education levels are depicted in Figure 2.7 and Figure 2.8. On top of each bar, the group total for the different ethnicity types are depicted. Based on the similarities within these figures, for further analysis the three biggest non-Dutch groups are combined (Moroccan, Surinamese, Dutch Antillean and Aruban and Turkish) as well as the ”From economies in transition” and ”Other” ethnicity groups, resulting in a total of four different ethnicity groups.

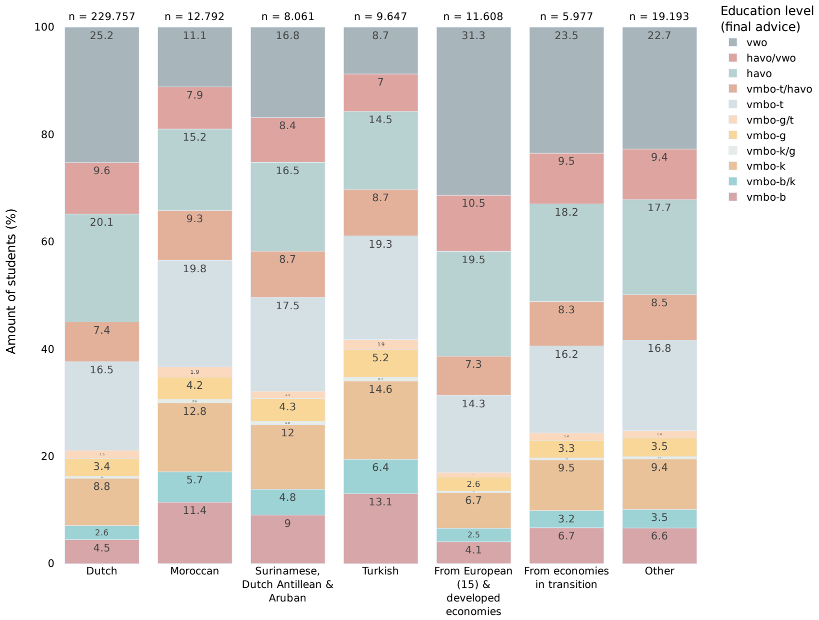


Figure 2.7 The percentage of students for the different education level of final advices grouped by ethnicity type. Group totals are depicted on top of each bar.

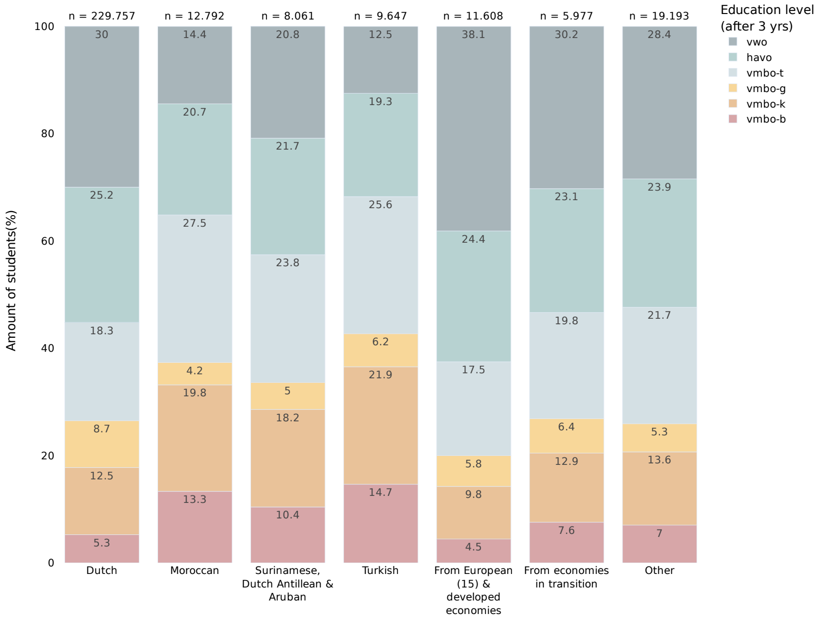


Figure 2.8 The percentage of students for the different education levels after 3 years grouped by ethnicity type. Group totals are depicted on top of each bar.

2.3.3 models and counterfactuals

The results on the CEVAE approximate inference method were examined by looking at the different loss components of the model during training (see Figure 2.9). The losses on the training batch are calculated for each iteration and on the test set each 100 iterations. In the upper graphs, the reconstruction loss for the different attributes of the model are plotted: l, s (separately for the continuous and binary part, denoted as s_con and s_bin respectively), t, e, f and y. The graph on the bottom right displays the regularization loss of the CEVAE. The combined loss of the reconstruction and regularization loss is depicted in the bottom left graph.

The counterfactual fairness of the selection procedure is evaluated by investigating the

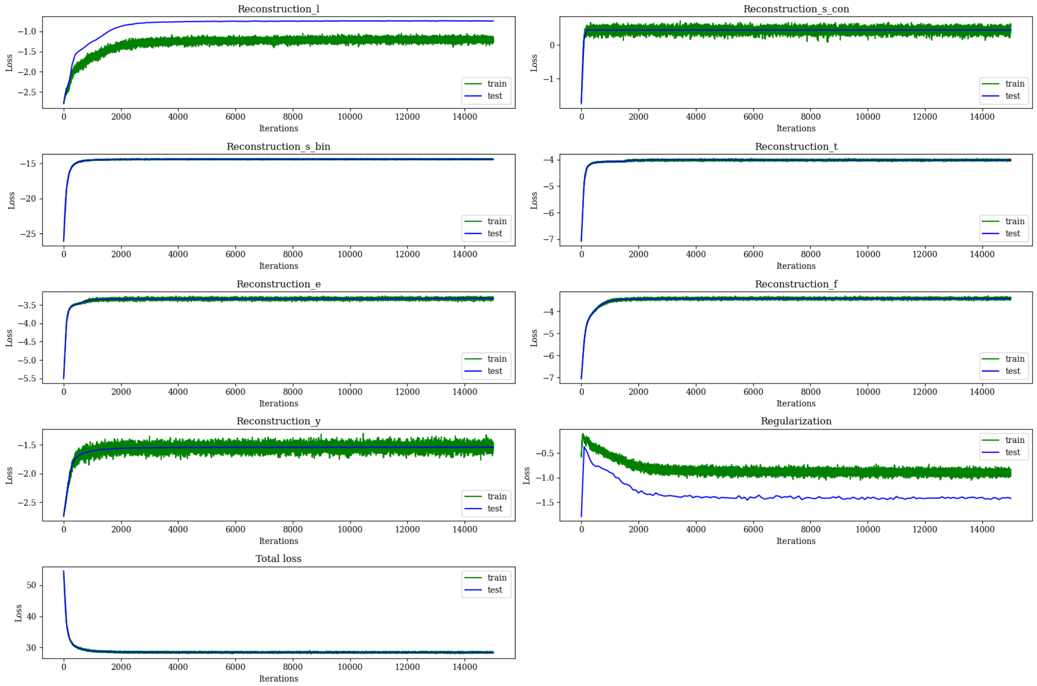


Figure 2.9 Loss components of the CEVAE. In green the loss on the training set and in blue the loss on the test set. The first seven graphs depict the reconstruction losses of the attributes l, s_con, s_bin, t, e, f and y. The bottom right shows the regularization loss and the bottom left the combined loss.

difference in outcome for real data $\hat{Y}$ and the outcome for generated counterfactual data $\hat{Y}_{do(A=a')}$. The counterfactual data are generated by altering the value of the sensitive attribute from *Dutch* → *Moroccan, Surinamese, Dutch Antillean and Aruban and Turkish*, from *Dutch* → *From EU-15 and developed countries*, from *Dutch* → *Other* and altering the other way around. Figure 2.10 shows the most important result of the approach of altering the value from *Dutch* → *Moroccan, Surinamese, Dutch Antillean and Aruban and Turkish*. The mean difference in outcome by the generation of these counterfactual datapoints was 1.02.

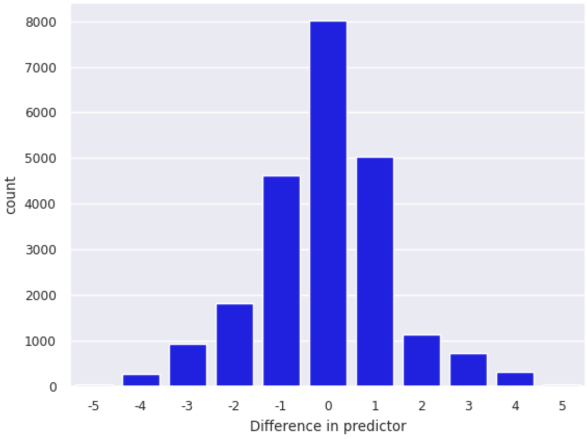


Figure 2.10 Difference between outcome $\hat{Y}$ and $\hat{Y}_{do(A=a')}$ by altering a from *Dutch* → *Moroccan, Surinamese, Dutch Antillean and Aruban and Turkish*. Positive numbers show a higher predicted education level in $\hat{Y}$ compared to $\hat{Y}_{do(A=a')}$.

The ML models were built with different input selections. The stability of the models was tested by doing ten repetitions. The models with the input $\{U\}$ had a mean accuracy of 65.0% ($\pm$0.2), a confidence of the model in the case of correct predictions of 67.4% ($\pm$0.2) and a confidence in incorrect cases of 59.5% ($\pm$0.2). The model with input $\{U, F\}$ had 85.7% ($\pm$0.1) accuracy, 84.0% ($\pm$0.2) confidence correct and 77.0% ($\pm$0.3) confidence incorrect. The fairness unawareness input $\{U, L, S, T, E, F\}$ gave an accuracy of 85.0% ($\pm$0.1), a confidence correct of 82.1% ($\pm$0.5) and a confidence incorrect of 77.7% ($\pm$0.5). The full model with input $\{U, A, L, S, T, E, F\}$ had an accuracy of 84.9% ($\pm$0.2), a confidence correct of 81.1% ($\pm$0.5) and a confidence incorrect of 77.0% ($\pm$0.6). See Table 2.2 for an overview of the results.

The performances of the different models are also displayed using confusion matrices, illustrated in Figure 2.12. Looking at the predicted and true class, one can have an idea about the type of mistakes of the model. Note that for the accuracies in Table 2.2, the highest two categories are selected, while for the confusion matrices we focused solely on the highest value.

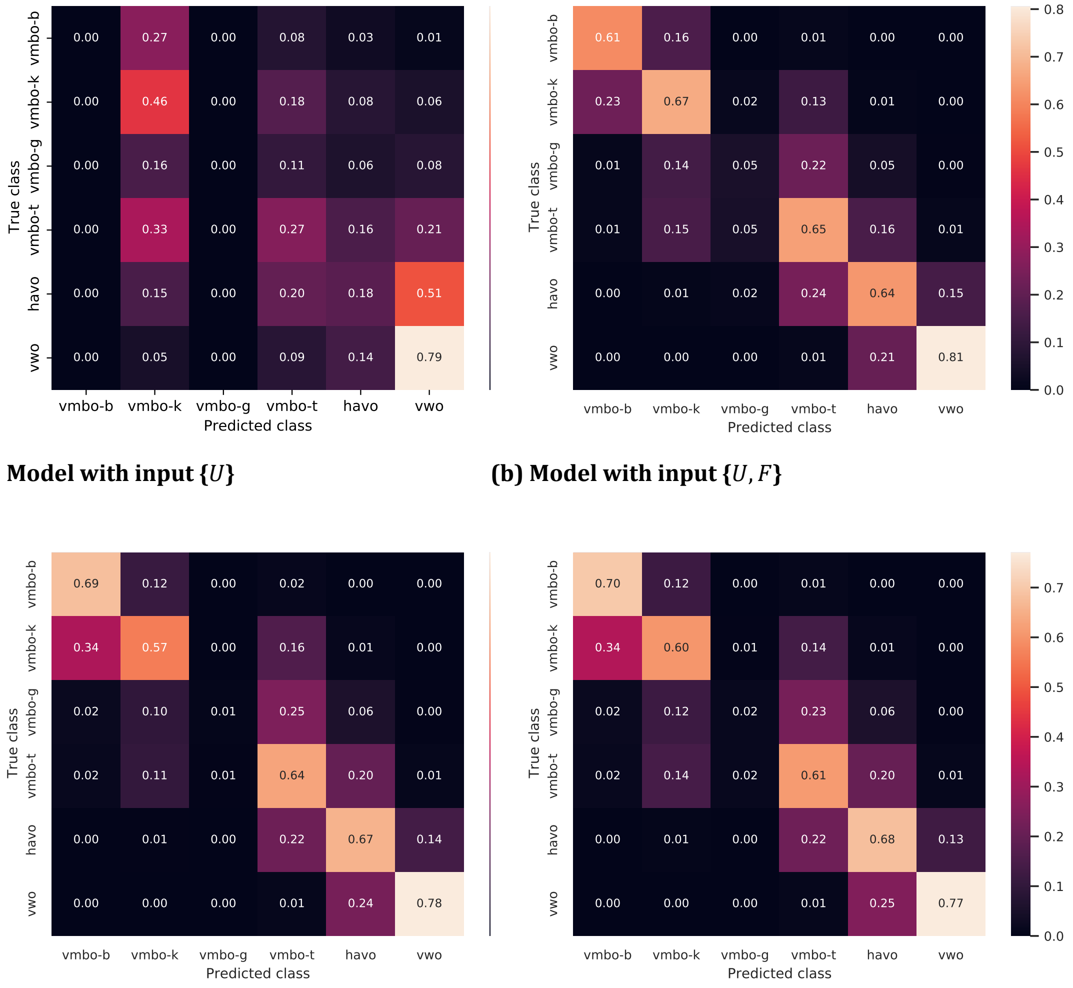
Table 2.2 Accuracies of the ML models with different input selections. The mean accuracy and standard deviation (SD) of ten repetitions is shown. *Confidence correct* shows the confidence of the model for correct predictions and the *confidence incorrect* shows the confidence of the model in incorrect cases.

Input	Mean accuracy ($\pm$SD) (%)	Confidence correct ($\pm$SD) (%)	Confidence incorrect ($\pm$SD) (%)
$\{U\}$	65.0 (0.2)	67.4 (0.2)	59.5 (0.2)
$\{U, F\}$	85.7 (0.1)	84.0 (0.2)	77.0 (0.3)
$\{U, L, S, T, E, F\}$	85.0 (0.1)	82.1 (0.5)	77.7 (0.5)
$\{U, A, L, S, T, E, F\}$	84.9 (0.2)	81.1 (0.5)	77.0 (0.6)

Fairness of the models was also tested using EOP as a fairness measure. For all models, the $P(\hat{Y} = 0|Y = 1)$ was measured, whereby 0 indicated a lower predicted educational attainment level than the true outcome. Here the ordering as described at the end of section 2.2.2 is used, which could alternatively be interpreted as a measure for the relative proportion of practical/vocational skills and theoretical/academic skills associated with these educational categories. See Table 2.3 for an overview of the proportion of pupils for the different models. The closer the EOP values of the different ethnicity groups, the more fair the algorithm, an equal value being the absolute fair scenario which is not the case. The model with input $\{U\}$ exposes the most similar EOP values.

Table 2.3 EOP of the ML models with different input selections. The EOP for the different ethnicity groups are displayed.

Input	EOP			
Input	Dutch (%)	Moroccan, Surinamese, Dutch Antillean and Aruban and Turkish (%)	From EU-15 or other similarly developed countries (%)	Others (%)
$\{U\}$	18.0	17.4	19.7	17.7
$\{U, F\}$	13.9	18.5	15.5	18.8
$\{U, L, S, T, E, F\}$	14.1	18.0	17.8	19.7
$\{U, A, L, S, T, E, F\}$	16.7	18.4	18.1	19.2



(a) Model with input $\{U\}$

(b) Model with input $\{U, F\}$

(c) Model with input $\{U, L, S, T, E, F\}$

(d) Model with input $\{U, A, L, S, T, E, F, U\}$

Figure 2.12 Confusion matrices of the ML models with different input variables. The x-axis represents the predicted class. The y-axis represents the true class. The numbers show the percentage of cases.

2.4 Discussion of case study

The first, preparatory, step of this research is to identify the factors that play a role in the decision making process for students' allocation to secondary education levels, in order to construct a valid causal model. This causal model allows systematic exploration of possible biases in the selection procedure, some of which the current literature also indicates to be present. The most important identified factors included the language level of pupils, SES with interrelated factors such as parental education level and bias in the teachers' estimates.

It is important to emphasise that the primary intention of the case study is to present and develop the technique, rather than to judge in an absolute sense the (un)fairness of selection processes in the Dutch educational system. There are several reasons why such an interpretation would be incautious, the most important of which is that the full causal graph is pared down.

With the pared down causal graph as adopted for the study, the counterfactual fairness can be determined, which is the second step of the research. Relationships in the causal graph, as well as the unknown variable intelligence, are estimated by maximizing the variational lower bound of the CEVAE model. By sampling from the learned distributions, the effect of generating counterfactual data, i.e. altering the value of the sensitive attribute, are examined.

The method reveals a slight decrease in predicted education level when altering the sensitive attribute from *Dutch → Moroccan, Surinamese, Dutch Antillean and Aruban and Turkish*. In addition, to propose fair ML models, models with different input selections are built. The model with input selection U (i.e. only the latent variable intelligence) yielded an accuracy of around 65%, while the models with more input variables all yielded an accuracy of around 85%. For all models, the confidence of the model was lower on average in the case of an incorrect prediction compared to a correct prediction. This can be considered as extra information that can contribute to the performance of the models. Confusion matrices revealed a high-level estimation of actual class or one of the next classes. ”Vmbo-g” was difficult to predict for all models. Specifically the model with only $\{U\}$ as input showed a low accuracy for ”vmbo-b” and ”vmbo-g”, the two smallest categories, depicted in Figure 2.8. The model with input selection $\{U\}$ did score the highest on the EOP fairness metric, albeit its relatively low accuracy. The revealed inequalities within the current data sample by inspecting the data are similar to previous studies (Lek, 2020). In line with the previous research results, testing the CEVAE with counterfactual data showed some effect of altering the sensitive attribute. This confirms the finding that the selection process, at least in the pared-down version of its causal graph analysed here, contains bias. For the subsequent analysis, when the only ”counterfactually fair” attribute was taken into account for the model (the non-descendant of sensitive attributes) the model yielded a lower accuracy than the models that considered more input variables. This is in line with previous work (Helwegen et al., 2020; Kusner et al., 2018) and suggests that the model has to give up some accuracy for the predictor to be more fair. Another way of describing this would be to state that a fair prediction can apparently not be as definitive or absolute in its pronouncements. In particular, as the data inherits the prejudices of prior decision makers, Y is also biased.

Counterfactual fairness as a measure does not require high accuracy of the model *per se*, by definition. What is surprising is the relatively big change in accuracy compared to previous research (Kusner et al., 2018). A possible explanation for this might be that only U can be selected as attribute or that the current method is not ideal. Perhaps none of the data available in a primary school setting are sufficiently informative for deciding which is the most appropriate vocational training, leading to particularly poor results for assessing which of the several possible ”vmbo” classes is most appropriate for any given pupil.

The results of this study should be interpreted with caution, given that it is based on a pared-down causal graph. Even though a lot of literature studies are incorporated in the construction of it, the causal graph still carries a lot of assumptions. A further extended analysis would be required to examine every causal link in this graph to asses the objective support for its existence, which holds both for the full causal graph and the pared-down version of it. This is especially of concern when counterfactuals are tested on a graph. This is especially the case because in this study the evidence for causal relationships is analysed using approximate inference instead of exact inference, for ease of computation and also for inferring unobserved variables. Besides, the way the unknown variable U (intelligence) is inferred is the inverse of what happens in real life, where U has a causal effect on the attributes L, T, E F and Y. When assuming the causal graph, we assume U is an exogenous variable that can only be explained by effects outside of the model. However, in this study, since it is believed that U has an effect on all its child nodes, and U is restricted to be of a lower dimension, U is believed to hold high dimensional features of L, T, E F and Y. Another note should be taken of the fact that this study was not able to use Y for the inference of U for out-of-sample prediction purposes. Some information for inferring U is lost this way. Yet, given these shortcomings, some information is stored in the variable U as indicated by the performance of the model with as only input U.

The findings of this study are also limited by the possible selection bias that occurred by dropping pupils that contained missing data (see Footnote 1). For example, it is known that for immigrants less data on highest attained education is available at CBS (CBS, 2021). In addition, CBS does not possess data on non-subsidised education and education taken abroad. Selection bias occurs when the sample is not a good representative of the whole population and although CBS has records on all pupils in the Netherlands, only a part of the data is used for the analysis. Besides, the models were less accurate on smaller groups, for example in pupils that ended up in ”vmbo-b” and ”vmbo-g”, due to sample size disparity. The educational level ”vmbo” was divided into the four subgroups to include all possibilities and to resemble reality. For future research, it would be interesting to see if combining them would solve the problem. During the testing of the models, it became apparent that all models were worse at predicting pupils from the ethnicity groups *Moroccan, Surinamese, Dutch Antillean and Aruban and Turkish*. This result may be explained by the fact that the outcome for these groups is the most distinctive. The same can happen for minority groups in decision making problems in the real world (Sandoval and Miille, 1980).

3 Conclusions

The aim of this research is to explore how causality inference can play a role in improving the evidence base for evidence-based governance. In essence the use of causality inference would be the next step. Beyond merely identifying correlations and trends, which are currently routine activities for national statistical institutes, it would be worthwhile to explore what evidence there is for those correlations to be causal in character. Given the context that national statistical institutes by their very nature work in, controlled, isolated experiments are not possible to perform. At least there is no feasible route to do so by design, although in some cases ’natural experiments’ in the sense of Imbens and Rubin (2015) may be available. The availability at CBS of many good quality time series can be of some help to assess Granger or predicitive causality, as is demonstrated in section 1, but does require a solid foundation in the sense that **all** relevant time series ought to be taken into account. A systematic process for identifying which variables are required is through the construction of causal graphs.

In the design of such causal graphs, and testing this design for correctness, another form of causality, *counterfactual causality* can play a role especially if large cross-sectional datasets are available so that the multivariate data space is covered well. The techniques of counterfactual analysis are becoming more prevalent, see e.g. Piccialli et al. (2023) for a very recent example in machine learning. For the case study presented here a paricular form of that, *counterfactual fairness*, is employed to explore potential biases in the decision process which leads to school level advice.

It bears repeating from the concluding paragraph of sect. 2.2.1 that the analysis in the paper is exploratory. The primary purpose here is to demonstrate the technique and therefore a choice is made to work with a pared-down version of the full causal diagram. A more complete analysis would certainly appear to be worthwhile but is beyond the scope of this research. The actual (un)fairness of the Dutch education system should only be judged on the basis of a more complete causal model and not on the pared-down version analysed here.

The case study on fairness demonstrates that counterfactuals can be used to get an estimation of fairness, which is therefore a viable alternative as it is not possible to do interventions on the real world system to get answers on fairness questions. In general, it seems that predicting the educational level of pupils three years ahead is a difficult task. This study indicates that including additional attributes for predicting the outcome (besides intelligence), increased the accuracy but lowered the fairness of the model. These extra attributes that appear to be good predictiors of the educational level, might also be considered unfair, like socio-economic status (Timmermans et al., 2018) or situation at home (Bijl, 2020), but this also depends on normative notions of fairness. These results, and also the indirect indications from e.g. the amount of switching between school levels of pupils, indicate that estimating future education at the age of 12 might be too early.

Further work is required to replicate and validate this specific counterfactual fairness application. Besides, multiple causal models can be put forward and tested. Regarding fairness in the education domain, while this study offered multiple ML models with different input selections, future research could identify fair pathways to ultimately construct ML models. Experts can be interviewed for this purpose, which was outside the scope of the current study. The location of schools can also be considered in future research, as the bias might differ between cities and villages (Vogels et al., 2021). Especially in cities with *”superdiversity”* like in Amsterdam, where 63% of the pupils have a non-Dutch ethnicity (Crul, 2016), bias in the selection procedure remains an important issue.

Acknowledgements : The authors would like to thank Marijke Hartgers and Karolijne van der Houwen for their invaluable input and expertise in the field of Dutch education statistics and the various pitfalls and selection effects that need to be taken into account when using the CBS data for the application reported here. Also, Reinoud Stoel is thanked for his careful reading and thoughtful suggestions which improved the manuscript.
The views expressed in this report are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Appendix

I Dutch Laws on Equality and Education

Article 1 of the Dutch constitution states equal treatment of individuals in equal circumstances and prohibition of discrimination of individuals on the grounds of religion, belief, political opinion, race or sex or any other grounds (Nederlandse Overheid, 1815):

Art. 1
Allen die zich in Nederland bevinden, worden in gelijke gevallen gelijk behandeld. Discriminatie wegens godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht of op welke grond dan ook, is niet toegestaan.

The law is further elaborated on in the Algemene wet gelijke behandeling (Equal Treatment Act) (Nederlandse Overheid, 1994).
Article 23 of the Dutch constitution states that the government is responsible for education (both private and public), that everyone is free to give education and that the government should monitor the quality of education (Nederlandse Overheid, 1815). Besides, children in the Netherlands are obligated to follow education by the Leerplichtwet 1969 (Compulsory Education Act) (Nederlandse Overheid, 1968).
In addition, Article 28 and Article 29 of the United Nations Convention on the Rights of the Child specify the right of education for each child and optimal development of the child's potential, respectively (United Nations, 1989):

Article 28
1. States Parties recognize the right of the child to education, and with a view to achieving this right progressively and on the basis of equal opportunity, they shall, in particular:
(a) Make primary education compulsory and available free to all;
(b) Encourage the development of different forms of secondary education, including general and vocational education, make them available and accessible to every child, and take appropriate measures such as the introduction of free education and offering financial assistance in case of need;
(c) Make higher education accessible to all on the basis of capacity by every appropriate means;
(d) Make educational and vocational information and guidance available and accessible to all children;
(e) Take measures to encourage regular attendance at schools and the reduction of drop-out rates.
2. States Parties shall take all appropriate measures to ensure that school discipline is administered in a manner consistent with the child's human dignity and in conformity with the present Convention.
3. States Parties shall promote and encourage international cooperation in matters relating to education, in particular with a view to contributing to the elimination of ignorance and illiteracy throughout the world and facilitating access to scientific and technical knowledge and modern teaching methods. In this regard, particular account shall be taken of the needs of developing countries.
Article 29
1. States Parties agree that the education of the child shall be directed to:
(a) The development of the child's personality, talents and mental and physical abilities to their fullest potential;

(b) The development of respect for human rights and fundamental freedoms, and for the principles enshrined in the Charter of the United Nations;
(c) The development of respect for the child's parents, his or her own cultural identity, language and values, for the national values of the country in which the child is living, the country from which he or she may originate, and for civilizations different from his or her own;
(d) The preparation of the child for responsible life in a free society, in the spirit of understanding, peace, tolerance, equality of sexes, and friendship among all peoples, ethnic, national and religious groups and persons of indigenous origin;
(e) The development of respect for the natural environment.
2. No part of the present article or article 28 shall be construed so as to interfere with the liberty of individuals and bodies to establish and direct educational institutions, subject always to the observance of the principle set forth in paragraph 1 of the present article and to the requirements that the education given in such institutions shall conform to such minimum standards as may be laid down by the State.

References

Aalders, P., A. M. L. van Langen, K. Smits, D. van den Tillaart, and M. H. J. Wolbers (2020). Pisa-2018 de verdieping: Kansenongelijkheid in het voortgezet onderwijs. *Nijmegen: KBA Nijmegen*.

Agirdag, O., P. Van Avermaet, and M. Van Houtte (2013). School segregation and math achievement: A mixed-method study on the role of self-fulfilling prophecies. *Teachers College Record 115(3)*, 1–50.

Barocas, S., M. Hardt, and A. Narayanan (2019). *Fairness and Machine Learning*. fairmlbook.org.

Bickel, P., A. Hammel, and J. O'Connell (1975, 2). Sex bias in graduate admissions: Data from berkeley. *Science 187*, 398–404.

Bijl, H. (2020). Parool: ”thuissituatie weegt niet meer mee bij schooladvies achtstegroepers”. Accessed on 12/09/2021.

Bisschop, P., E. van den Berg, K. van der Ven, W. de Geus, and D. Kooij (2019). Aanvullend en particulier onderwijs. onderzoek naar de verschijningsvormen en omvang van aanvullend en particulier onderwijs en motieven voor deelname. in opdracht van het ministerie van onderwijs, cultuur en wetenschap, oberon en seo economisch onderzoek.

Bol, T. (2020). Inequality in homeschooling during the corona crisis in the netherlands. first results from the liss panel. Accessed on 23/07/2021.

Boone, S. and M. Van Houtte (2013). Why are teacher recommendations at the transition from primary to secondary education socially biased? a mixed-methods research. *British Journal of Sociology of Education 34(1)*, 20–38.

Boterman, W. R. (2019). The role of geography in school segregation in the free parental choice context of dutch cities. *Urban Studies 56(15)*, 3074–3094.

CBS (2018). Sociaaleconomische positie. https://www.cbs.nl/nl-nl/achtergrond/2018/47/sociaaleconomische-positie, Accessed on 31/05/2021.

CBS (2021). Hoe verschillen opleiding en schoolkeuze naar migratieachtergrond? https://www.cbs.nl/nl-nl/dossier/dossier-asiel-migratie-en-integratie/hoe-verschillen-opleiding-en-schoolkeuze-naar-migratieachtergrond-, Accessed on 31/05/2021.

Chung, J., K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio (2015). A recurrent latent variable model for sequential data. *Advances in neural information processing systems 28*, 2980–2988.

Clevert, D. A., T. Unterthiner, and S. Hochreiter (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 1–14.

Crul, M. (2016). Super-diversity vs. assimilation: how complex diversity in majority–minority cities challenges the assumptions of assimilation. *Journal of Ethnic and Migration Studies 42(1)*, 54—68.

Denessen, E. J. P. G. (2017). Verantwoord omgaan met verschillen: sociale-culturele achtergronden en differentiatie in het onderwijs. Doctoral dissertation, Leiden.

Dijks, M. A., M. J. Warrens, E. Fleur, H. Korpershoek, I. J. M. Wichgers, and R. J. Bosker (2020). The predictive power of track recommendations in dutch secondary education. *Pedagogische studien 97(4)*, 263–280.

Ditton, H., J. Krüsken, and S. M. (2005). Bildungsungleichheit - der beitrag von familie und schule. *Zeitschrift Für Erziehungwissenschaft 8*, 285—304.

Dumont, H., D. Klinge, and K. Maaz (2019). The many (subtle) ways parents game the system: Mixed-method evidence on the transition into secondary-school tracks in germany. *Sociology of Education 92 (2)*, 199–228.

Engle, R. and C. Granger (1987, 3). Co-integration and error correction: Representation, estimation, and testing. *Econometrica 55*(2), 251–276.

Gentrup, S., G. Lorenz, C. Kristen, and I. Kogan (2020). Self-fulfilling prophecies in the classroom: Teacher expectations, teacher feedback and student achievement. *Learning and Instruction 66*, 101296.

Geven, S., A. Batruch, and H. van de Werfhorst (2018). Inequality in teacher judgements, expectations and track recommendations: A review study. Universiteit van Amsterdam: Amsterdam Institute for Social Science Research (AISSR).

Glorot, X., A. Bordes, and Y. Bengio (2011). Deep sparse rectifier neural networks. *Proceedings of the fourteenth international conference on artificial intelligence and statistics 15*, 315—323.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press.

Granger, C. (1969, 8). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*(3), 424–438.

Gregor, K., I. Danihelka, A. Graves, D. Rezende, and D. Wierstra (2015). Draw: A recurrent neural network for image generation. *International Conference on Machine Learning 37*, 1462–1471.

Hacking, I. (2006). Making up people. *London Review of Books 28*(16), 1.

Han, J. and C. Moraga (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. *Lecture Notes in Computer Science 930*, 195—201. International workshop on artificial neural networks.

Hausman, D. M. and J. Woodward (1999). Independence, invariance and the causal markov condition. *The British journal for the philosophy of science 50(4)*, 521–583.

Helwegen, R., C. Louizos, and P. Forré (2020). Improving fair predictions using variational inference in causal models. *arXiv cs.LG/2008.10880*.

Hop, M. and H. van Boxtel (2013). Wetenschappelijke verantwoording cito intelligentietest vo. Accessed on 02/06/2021.

Imbens, G. and D. Rubin (2015). *Causal inference for statistics, social, and biomedical sciences. An introduction*. Cambridge university press.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv cs.LG/1412.6980*.

Kingma, P. and M. Welling (2013). Auto-encoding variational bayes. *arXiv cs.LG/1312.6114*.

Kusner, M., J. Loftus, C. Russell, and R. Silva (2018). Counterfactual fairness. *arXiv:1703.06856*.

Kuyper, H. and G. Van der Werf (2012). Excellente leerlingen in het voortgezet onderwijs. schoolloopbanen, risicofactoren en keuzen.

Lek, K. M. (2020). Teacher knows best? on the (dis) advantages of teacher judgments and test results, and how to optimally combine them. *Doctoral dissertation, Utrecht University*.

Louizos, C., U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling (2017). Causal effect inference with deep latent-variable models. *arXiv cs.LG/1705.08821*.

Louizos, C., K. Swersky, Y. Li, M. Welling, and R. Zemel (2015). The variational fair autoencoder. *arXiv cs.LG/1511.00830*.

Marioni, R. E., G. Davies, C. Hayward, D. Liewald, S. M. Kerr, A. Campbell, M. Luciano, B. H. Smith, S. Padmanabhan, L. J. Hocking, N. D. Hastie, A. F. Wright, D. J. Porteous, P. M. Visscher, and I. J. Deary (2014). Molecular genetic contributions to socioeconomic status and intelligence. *Intelligence 44*, 26–32.

Min. OCW (2017). Subsidieregeling gelijke kansen in het onderwijs. https://wetten.overheid.nl/jci1.3:c:BWBR0040077/2017-10-19, Accessed on 15/11/2021.

Ministerie van Onderwijs Cultuur en Wetenschap (2021). Schooladvies en heroverweging schooladvies. https://www.onderwijsincijfers.nl/kengetallen/po/leerlingen-po/prestaties-schooladvies, Accessed on 12/05/2021.

Nederlandse Overheid (1815). Grondwet. https://wetten.overheid.nl/jci1.3:c:BWBR0001840/2018-12-21, Accessed on 13/11/2021.

Nederlandse Overheid (1968). Leerplichtwet 1969. https://wetten.overheid.nl/jci1.3:c:BWBR0002628/2021-10-01, Accessed on 13/11/2021.

Nederlandse Overheid (1994). Algemene wet gelijke behandeling. https://wetten.overheid.nl/jci1.3:c:BWBR0006502/2020-01-01, Accessed on 13/11/2021.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J., M. Glymour, and N. P. Jewell (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Peters, J., D. Janzing, and B. Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Piccialli, V., D. R. Morales, and C. Salvatore (2023). Supervised feature compression based on counterfactual analysis. *arXiv cs.LG/2211.09894v3*.

Pijpers, F. and I. Wanders (1994). Reverberation mapping of active galactic nuclei: the sola method for time-series inversion. *Mon.Not.Roy.Ast.Soc. 271*, 183–196.

Pijpers, F. P. (2021). A non-parametric method for determining epidemiological reproduction numbers. *Journal of Mathematical Biology 82*(5), 37.

PO-raad (2019). Extra ondersteuning moet voor alle leerlingen toegankelijk zijn. https://www.poraad.nl/nieuws-en-achtergronden/extra-ondersteuning-moet-voor-alle-leerlingen-toegankelijk-zijn, Accessed on 24/07/2021.

Rezende, D., S. Mohamed, and D. Wierstra (2014). Stochastic backpropagation and approximate inference in deep generative models. *International conference on Machine Learning*, 1278–1286.

Sandoval, J. and M. W. Miille (1980). Accuracy of judgments of wisc-r item difficulty for minority groups. *Journal of Consulting and Clinical Psychology 48(2)*, 249–253.

Schwab, P., L. Linhardt, and W. Karlen (2018). Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*.

Shalit, U., F. D. Johansson, and D. Sontag (2017). Estimating individual treatment effect: generalization bounds and algorithms. *International Conference on Machine Learning*, 3076–3085.

Shavit, Y. and H. P. Blossfeld (1993). Persistent inequality: Changing educational attainment in thirteen countries. *San Fransisco: Westview Press*.

Spinath, B., C. Eckert, and R. Steinmayr (2014). Gender differences in school success: What are the roles of students' intelligence, personality and motivation? *Educational Research 56(2)*, 230–243.

Spirtes, P., C. N. Glymour, R. Scheines, and D. Heckerman (1993). *Causation, prediction, and search*. Springer.

Tieben, N. (2011). Parental resources and relative risk aversion in intra-secondary transitions: A trend analysis of non-standard educational decision situations in the netherlands. *European Sociological Review 27(1)*, 310–42.

Tieleman, T. and G. Hinton (2014). Rmsprop gradient optimization. Accessed on 05/11/2021.

Timmermans, A., H. Kuyper, and G. van der Werf (2013). *Schooladviezen en onderwijsloopbanen*. GION, Gronings Instituut voor Onderzoek van het Onderwijs.

Timmermans, A. C., H. De Boer, H. T. A. Amsing, and M. P. C. Van Der Werf (2018). Track recommendation bias: Gender, migration background and ses bias over a 20-year period in the dutch context. *British Educational Research Journal 44 (5)*, 847—874.

Timmermans, A. C., H. Kuyper, and G. Van der Werf (2015). Accurate, inaccurate, or biased teacher expectations: Do dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology 85*, 459–478.

Tolsma, J., M. Coenders, and M. Lubbers (2007). Trends in ethnic educational inequalities in the netherlands: a cohort design. *European Sociological Review 23(3)*, 325–339.

United Nations (1989). Convention on the rights of the child. https://www.ohchr.org/EN/ProfessionalInterest/Pages/CRC.aspx, Accessed on 13/11/2021.

U.S.C. (1996). Texas house bill 588. https://capitol.texas.gov/billlookup/text.aspx?LegSess=75R&Bill=HB588, Accessed on 13/11/2021.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Van de Werfhorst, H. and J. Mijs (2010). Achievement inequality and the institutional structure of educational systems: a comparative perspective. *Annual Review of Sociology 36*, 407–428.

Van de Werfhorst, H. G. (2011). Selectie en differentiatie in het nederlandse onderwijsbestel: gelijkheid, burgerschap en onderwijsexpansie in vergelijkend perspectief. *Pedagogische studiën 88*.

Verma, S. and J. Rubin (2018). Fairness definitions explained. *2018 ieee/acm international workshop on software fairness (fairware)*, 1–7.

Vogels, R., M. Turkenburg, and L. Herweijer (2021). Samen of gescheiden naar school: De betekenis van sociale scheiding en ontmoeting. *Den Haag: Sociaal Cultureel Planbureau*.

Werfhorst, H. G. v. d. (2007). Onderwijsinstituties in nederland: gelijkheid, efficiëntie, allocatie en burgerschap? *J.W. Duyvendak & M. Otto (red.), Sociale kaart van Nederland. Over maatschappelijke instituties*, 133–151.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information