

Anleitung WSL-kNN

Einleitung

WSL-kNN ist ein Tool, das es ermöglicht, aufgrund bestehender Datensätze von Holzschlägen aus der Vergangenheit den Aufwand für einen zukünftigen Holzschlag zu schätzen. Dabei werden aufgrund bekannter Parameter des zukünftigen Holzschlags (z.B. Rückedistanz, Nadelholzanteil oder Vorrat im entsprechenden Gebiet) ähnliche Datensätze aus der Vergangenheit gesucht und gemittelt.

Das Tool WSL-kNN ersetzt das bisherige KnnWorkbook, welches auf Excel 2007 bzw. Excel 2010 basierte. Das neue Tool basiert auf der Programmiersprache Java und ist somit unabhängig von einer Excel-Installation nutzbar und auch auf verschiedenen Plattformen (Windows, Mac) lauffähig. Einzige Voraussetzung für den Betrieb des Tools ist die Installation einer Java Runtime Environment (JRE), welche, falls sie nicht bereits auf dem Computer installiert ist, auf www.java.com heruntergeladen werden kann.

Methode

Die Methode der k nächsten Nachbarn

Die Methode der k nächsten Nachbarn (kNN-Methode) ist eine Vorgehensweise der Statistik, bei welcher aus einer Menge bereits erhobener Datensätze bestimmte Merkmale eines Referenzdatensatzes geschätzt werden sollen. Die kNN-Methode beruht auf der Idee der Ähnlichkeit zweier Datensätze: Zwei Datensätze sind dann ähnlich, wenn ihre Merkmale ähnlich sind, wenn also je ein Merkmalswert des einen Datensatzes verhältnismässig nahe beim Merkmalswert des anderen Datensatzes liegt, und das für alle Merkmale gilt.

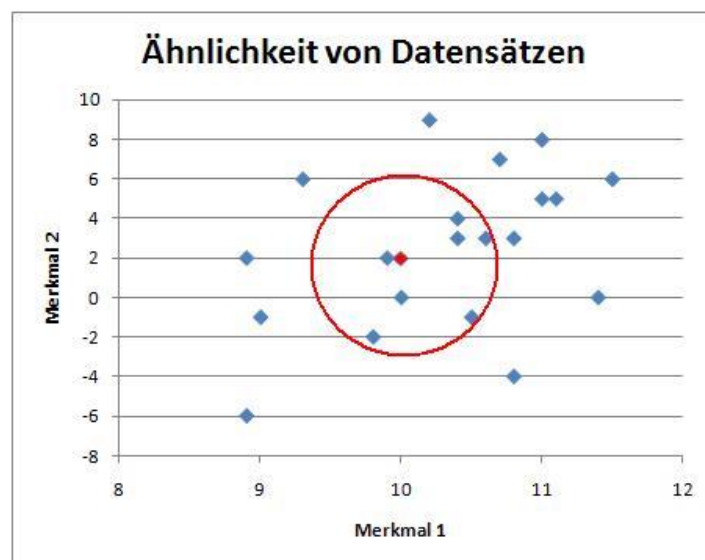


Abbildung 1: Beispiel von Datensätzen mit zwei Merkmalen.

Ein Beispiel dazu ist in Abbildung 1 ersichtlich. In einer Menge von 21 Datensätzen mit zwei Merkmalen (aufgetragen auf die x- und die y-Achse) sollen alle Datensätze gefunden werden, die innerhalb einer bestimmten Distanz zu einem ausgewählten Datensatz - dem **Referenzdatensatz** - liegen. In unserem Beispiel sind demnach 7 Datensätze ähnlich zum Referenzdatensatz. Wir bezeichnen solche ähnlichen Datensätze als **nächste Nachbarn**.

Natürlich kann das Mass der Ähnlichkeit nicht nur für zweidimensionale Datensätze sondern allgemein für m -dimensionale Datensätze bestimmt werden, wobei m der Anzahl von Merkmalen der Datensätze entspricht.

Der Parameter k

Die kNN-Methode wendet üblicherweise eine leicht unterschiedliche Strategie an. Anstatt einen maximalen Abstand festzulegen, innerhalb dessen sich die nächsten Nachbarn in Relation zum Referenzdatensatz befinden müssen, wird eine Anzahl nächster Nachbarn bestimmt, die danach weiter betrachtet werden. Diese Anzahl nächster Nachbarn wird durch den Parameter k ausgedrückt, für welchen im Normalfall eine nicht allzu grosse Ganzzahl eingesetzt wird, z.B. $k := 7$.

Bei dieser Vorgehensweise ist es hilfreich zu wissen, wie nahe die nächsten Nachbarn in Relation zum Referenzdatensatz sind. Liegen sie alle nahe beieinander? Dann wird die Schätzung wohl eine gute sein. Oder liegen sie weit entfernt? Dann wird die Schätzung wohl eher wenig Vertrauen erwecken.

Die kNN-Methode als Schätzmethode

Die kNN-Methode kann angewendet werden, um Schätzungen vorzunehmen. Mittels einer Reihe von **unabhängigen Merkmalen** soll ein einzelnes **abhängiges Merkmal** geschätzt werden. Wie beeinflussen beispielsweise die Rückedistanz, der Nadelholzanteil und das durchschnittliche Mittelstammvolumen die Kosten der Holzernte? Aus einer Reihe von bereits erhobenen Datensätzen sollen für einen zukünftigen Holzschlag diese Kosten abgeschätzt werden. Der zukünftige Holzschlag ist also der Referenzdatensatz, für welchen nur die Werte der unabhängigen Variablen a priori bekannt sind, nicht aber der Wert der abhängigen Variable. Aus den vorhandenen Datensätzen werden die k nächsten Nachbarn ausgesucht, und aus deren abhängigen Variablen wird ein Schätzwert für die abhängige Variable des Referenzdatensatzes berechnet. Dabei wird üblicherweise zusätzlich berücksichtigt, wie nahe/ähnlich jeder der nächsten Nachbarn dem Referenzdatensatz tatsächlich ist. Nahe nächste Nachbarn erhalten somit ein grösseres Gewicht bei der Schätzung als weiter entfernte nächste Nachbarn.

Euklidische Distanz

Um die Nähe eines Datensatzes zum Referenzdatensatz zu bestimmen wird oft die **euklidische Distanz** als Distanzmass eingesetzt. Die euklidische Distanz ist in Worten ausgedrückt *die Quadratwurzel aus der Summe der quadrierten Abstände zwischen den Merkmalen zweier Datensätze*. Bei der kNN-Methode werden die Abstände zusätzlich mit einem besonderen Faktor gewichtet, welcher die verschiedenen Einheiten und Skalen der Merkmale normiert. Die Formel lautet dann:

$$d'_{ij} = \sqrt{\sum_{p=1}^m \frac{\alpha_p^2}{\beta_p^2} (x_{ip} - x_{jp})^2}$$

d'_{ij} ist die euklidische Distanz zwischen einem Datensatz i und dem Referenzdatensatz j . \sum ist die Summe über alle unabhängigen Merkmale p . α_p ist der Korrelationskoeffizient zwischen dem unabhängigen Merkmal p und dem abhängigen Merkmal y . β_p ist die Standardabweichung des Merkmals p . x_{ip} und x_{jp} sind die jeweiligen Werte der unabhängigen Merkmale des Datensatzes i bzw. des Referenzdatensatzes j .

Auf [dx.doi.org/10.3188/szf.2012.0119](https://doi.org/10.3188/szf.2012.0119) ist ein Artikel verfügbar, welche die Methode im Detail beschreibt sowie ein Fallbeispiel mit der früheren Version dieses Tools zeigt.

Anwendung des Tools

Start des Tools

Nach dem Entpacken der Datei WslKnn.zip erhalten Sie die Datei WslKnn.jar, die per Doppelklick gestartet werden kann. Einzige Voraussetzung dafür ist, dass auf dem Computer eine aktuelle Java Version installiert ist (min. Java 8).

In gewissen Konstellationen kann es vorkommen, dass beim Doppelklick auf WslKnn.jar das Tool nicht gestartet wird, dafür jedoch ein Programm versucht, die Datei zu entpacken. Gehen Sie dann wie folgt vor: Erstellen Sie eine Datei mit dem Namen startWslKnn.bat, öffnen Sie diese mit einem einfachen Textbearbeitungsprogramm wie zum Beispiel Notepad, schreiben Sie «java -jar WslKnn.jar» (ohne Anführungs- und Schlusszeichen) in die Datei, und speichern Sie sie. Danach kann das Tool mit einem Doppelklick auf startWslKnn.bat gestartet werden.

Datenimport

Der erste Schritt für die Verwendung des Tools ist der Import von vorhandenen Datensätzen. Dies geschieht im Menü «Datei» über den Eintrag «CSV-Datei importieren». CSV steht für Comma Separated Value. Eine Datei in diesem Format erhält man beispielsweise, indem man eine vorhandene Excel-Tabelle öffnet und dann im Excel über «Speichern unter» das CSV-Format als gewünschtes Speicherformat wählt. Die Datei muss in der ersten Zeile die Überschriften enthalten, auf den nachfolgenden Zeilen kommen die Datensätze. Eine Beispieldatei befindet sich im zip-Archiv, mit welchem Sie dieses Tool erhalten haben.

Durch den Klick auf «CSV-Datei importieren» wird der Inhalt dieser Datei in die Haupttabelle geladen (Abbildung 2). Zusätzlich werden für alle Spalten, die numerische Werte erhalten, der Durchschnitt (arithmetisches Mittel) und die Standardabweichung angezeigt.

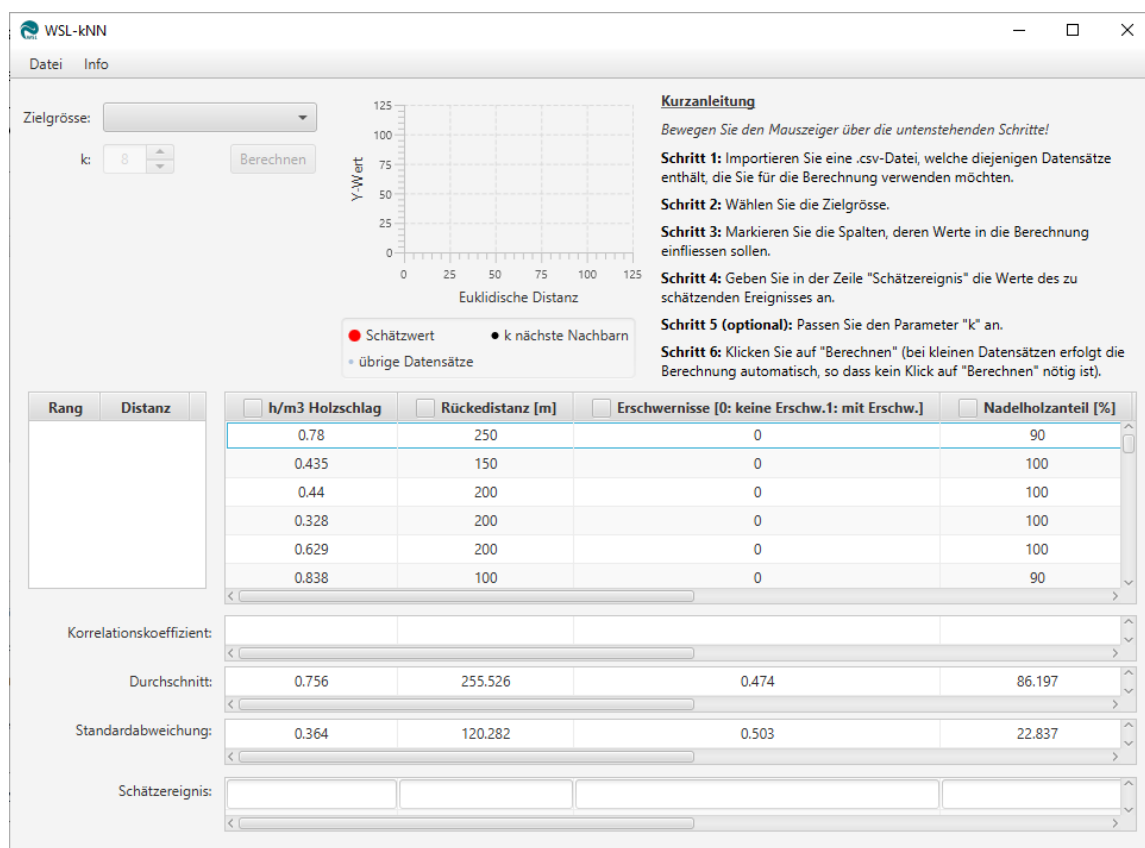


Abbildung 2: Ansicht des Tools nach dem importieren einer CSV-Datei.

Wahl der Zielgrösse

Als nächstes muss oben links die Zielgrösse ausgewählt werden. Die Zielgrösse ist diejenige Grösse, für welche ein Wert geschätzt werden soll, also in der Regel der zeitliche oder finanzielle Aufwand. Zur Verfügung stehen alle Spalten, die numerische Werte enthalten. Nach der Auswahl der Zielgrösse wird die gewählte Spalte grün hinterlegt. Ausserdem werden für alle Spalten die Korrelationskoeffizienten angezeigt, d.h. wie stark die jeweilige Spalte mit der Zielgrösse korreliert (Abbildung 3). Der Korrelationskoeffizient ist ein Wert zwischen -1 und 1. Je stärker entfernt der Korrelationskoeffizient von 0 ist, desto höher die Korrelation, und desto wertvoller ist die entsprechende Spalte für die Berechnung. Im Beispiel dieser Anleitung wurde die Spalte «h/m3 Holzschlag» gewählt, d.h. die Produktivität (Abbildung 3).

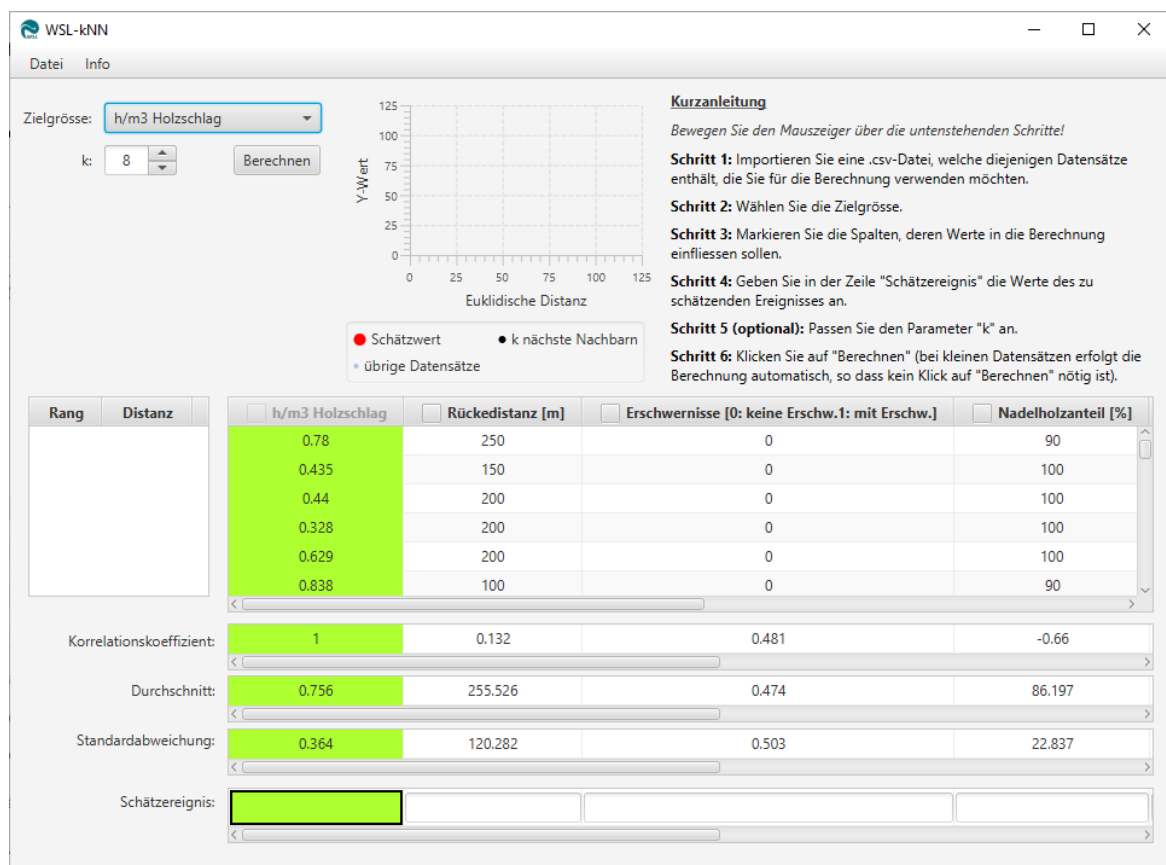


Abbildung 3: Ansicht des Tools nach dem Auswählen einer Zielgrösse.

Wahl der Merkmale und Eingabe der Referenzdaten

Als nächstes müssen diejenigen Spalten markiert werden, deren Werte in die Berechnung einfließen sollen. Dies geschieht durch das Setzen eines Häkchens in der entsprechenden Spaltenüberschrift. Sinnvollerweise werden diejenigen Spalten gewählt, welche am stärksten mit der Zielgröße korrelieren. Die gewählten Spalten werden blau hinterlegt (Abbildung 4). Die entsprechenden Felder in der Zeile «Schätzereignis» werden rot hinterlegt. Dort müssen nun Werte für alle selektierten Spalten eingegeben werden.

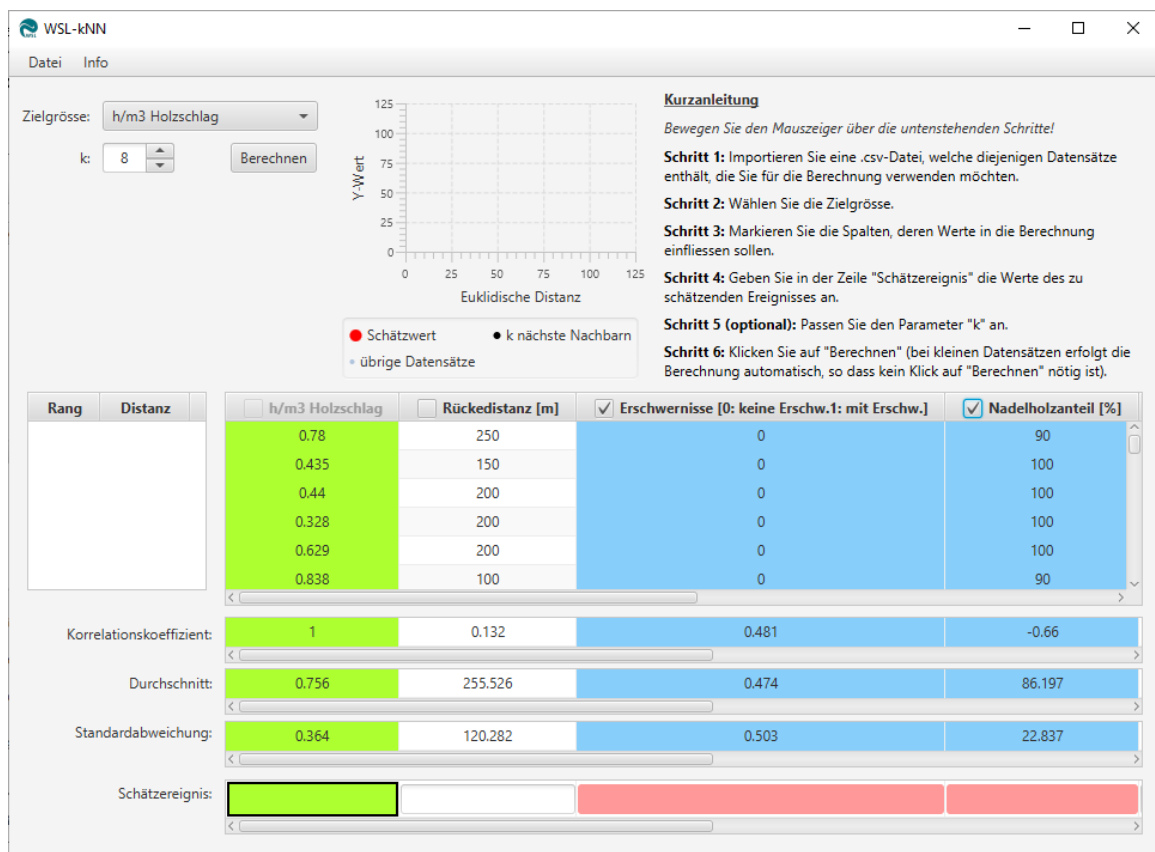


Abbildung 4: Ansicht des Tools nach Auswahl der zu berücksichtigenden Spalten.

Berechnung

Sobald in der Zeile «Schätzereignis» alle benötigten Werte eingegeben wurden, kann die Berechnung gestartet werden. Dies geschieht bei kleineren Datensätzen automatisch, noch während der Eingabe der Daten. Bei grösseren Datensätzen muss dazu auf den Button «Berechnen» geklickt werden. Danach wird in der Zeile «Schätzereignis» in der Spalte der gewählten Zielgrösse der berechnete Wert im grün hinterlegten, schwarz umrandeten Feld angezeigt (Abbildung 5). Zusätzlich werden links der Haupttabelle Rang und Distanz angezeigt. Die Distanz entspricht der euklidischen Distanz zwischen dem Datensatz in der entsprechenden Zeile und dem Schätzereignis. Der Datensatz mit der kürzesten Distanz zum Schätzereignis erhält Rang 1, der mit der zweitkürzesten Distanz Rang 2, etc. Fett markiert werden die k nächsten Nachbarn, d.h. die k Anzahl Datensätze mit dem tiefsten Rang, wobei der Parameter k oben links angepasst werden kann. Alle Tabellenspalten können auch sortiert werden, indem man in die Tabellenüberschrift der entsprechenden Spalte klickt (rechts vom Text).

Diese Distanz wird auch in der Grafik oberhalb der Haupttabelle graphisch dargestellt, zusammen mit der Zielgrösse. Anhand der Information in dieser Grafik kann der Parameter k angepasst werden, so dass er optimal auf die Daten abgestimmt ist. Ein Doppelklick auf das Diagramm öffnet eine vergrösserte Version des Diagramms in einem neuen Fenster.

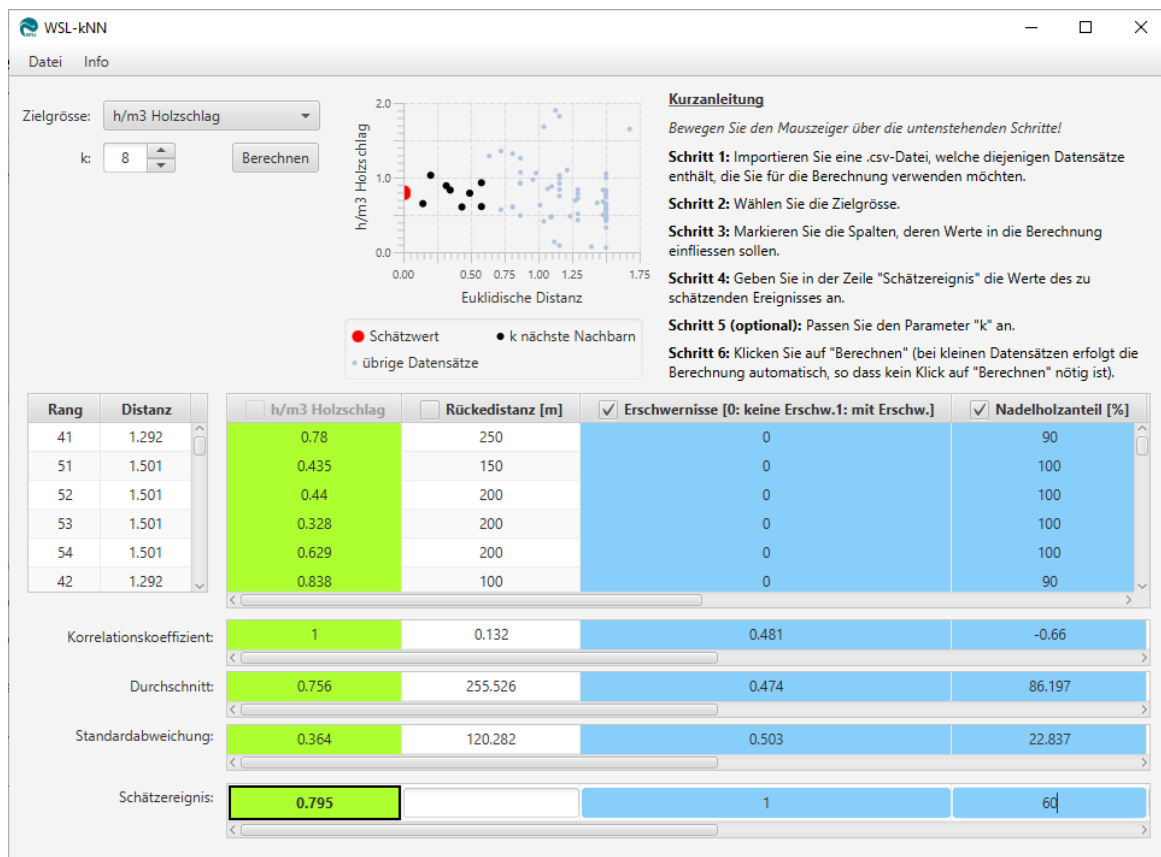


Abbildung 5: Ansicht des Tools nach dem Berechnungsvorgang.

Speichern und Laden von Projekten

Die gemachten Eingaben können über das Menü «Datei» gespeichert («Projekt speichern») bzw. geladen («Projekt laden») werden. Ein Projekt ist die Kombination der ursprünglich gewählten CSV-Datei und den weiteren gemachten Eingaben. Diese Informationen werden alle zusammen in einer Datei mit der Erweiterung knnx festgehalten.