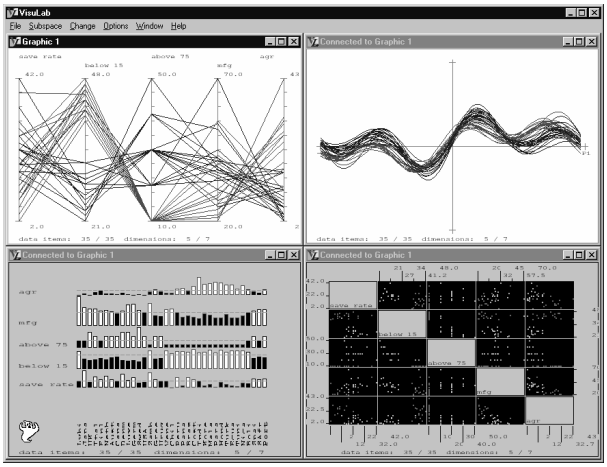
Visualisierung als Werkzeug zur Analyse mehrdimensionaler Daten

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Institut für Computational Science
Prof. Dr. H. Hinterberger

evim ETHz.ch

Praxismodul 3

Visualisierung als Werkzeug zur Analyse mehrdimensionaler Daten



Praxismodule

1 Publizieren über Internet – Frontpage
2 Tabellenkalkulation – Excel
3 Visualisierung – VisuLab
4 Daten verwalten (1) – Excel
5 Daten verwalten (2) – Access
6 Makroprogrammierung – Excel

Internetadresse: **www.evim.ethz.ch**

Wie bearbeite ich dieses Modul?

Dieses Praxismodul bearbeiten Sie am effizientesten, wenn Sie die folgenden drei Teile in angegebener Reihenfolge angehen:

Teil A: Einführung.....Seite 5

In diesem Praxismodul lernen Sie *mehrdimensionale Daten* zu visualisieren. Hier finden Sie eine kurze Einführung zum Thema Datenanalyse aus der Sicht der Anwender.

Teil B: *E.Tutorial*.....Seite 7

Das *E.Tutorial Praxis 3* ist ein computergestützter Lehrgang, der Ihnen in 13 Lektionen einen Einblick in Visualisierungsmethoden mit Hilfe von *Excel* und der speziellen Visualisierungssoftware *VisuLab* vermitteln soll.

In diesem *E.Tutorial* werden Sie...

- ...mehrdimensionale Daten in *Excel* und *VisuLab* visuell darstellen.
- ...mit *VisuLab* mehrdimensionale Daten analysieren.

In diesem Teil finden Sie zusätzlich eine Erinnerungshilfe (Reminder) für die einzelnen Lektionen des *E.Tutorials*, sowie einen Multiple-Choice Test.

Zeitaufwand: ca. 2 bis 3 Stunden

Teil C: Testaufgabe.....Seite 13

Bei der Testaufgabe werden Sie mit Hilfe der Visualisierungssoftware *VisuLab* eine Datensammlung von Irisblüten analysieren und damit eine Hypothese zur Abstammung einer Art überprüfen.

Zeitaufwand: ca. 1 Stunde

Begriffe:

In diesem Praxismodul werden folgende Begriffe behandelt:

Grafische Wahrnehmung
Visualisierungsmethode
Datensammlung
Datensatz
Dimensionalität von Daten
Häufigkeitsvergleich
Beziehungsvergleich
Erkundende Datenanalyse
Diagrammtyp
Parallel-Koordinaten
Punktediagramm-Matrix
Permutations-Matrix
Andrews' Kurven

Teil A: Einführung

Explorative Datenanalyse

Im letzten Praxismodul haben Sie gesehen, wie Sie mit Hilfe von Modellen und Simulationen (z.B. Populationswachstum von Arten, die in Nahrungs-Konkurrenz stehen) Rückschlüsse auf ein zu untersuchendes System machen können. Stehen Sie aber vor der Aufgabe, ein solches real existierendes System in einem Modell erstmals oder neu zu beschreiben, müssen Sie zuerst herausfinden, welches die relevanten Faktoren (Dimensionen) darstellen und wie diese miteinander zusammenhängen könnten. Dazu eignen sich Methoden der erkundenden (explorativen) Datenanalyse.

Ist Visualisieren eine wissenschaftliche Methode?

Die **visuelle Wahrnehmung** des Menschen ist stark darauf ausgerichtet, in Bildern Zusammenhänge zu erkennen. Manchmal werden wir dabei auch fehlgeleitet, wie z.B. bei optischen Täuschungen. Diese Fähigkeit unseres Gehirns kann auch dazu genutzt werden, um in grossen Datenmengen schnell Wichtiges von Unwichtigem zu unterscheiden. Da Ihre Daten meist als lange Zahlenkolonnen anfallen, die Sie unmöglich überblicken oder gar Zusammenhänge erkennen können, ist es von Vorteil, die Daten in eine visualisierte Form zu überführen.

Die Abbildung (**Abb. 1**) auf der nächsten Seite illustriert anhand eines Beispiels aus der Geschichte, wie wenig aussagende Zahlenkolonnen durch Visualisierung in eine Form gebracht werden können, in der Zusammenhänge (in Form von Geschichte) erkennbar werden. Darin ist Napoleons Russlandfeldzug 1812/13 dargestellt: Es zeigt die Bewegung der Armee (grau: Vormarsch, schwarz: Rückzug), die Anzahl der am Leben gebliebenen Soldaten (Breite der grauen und schwarzen Linie) und die Temperatur (Kurve in der unteren Hälfte des Diagramms) während des Rückzuges.

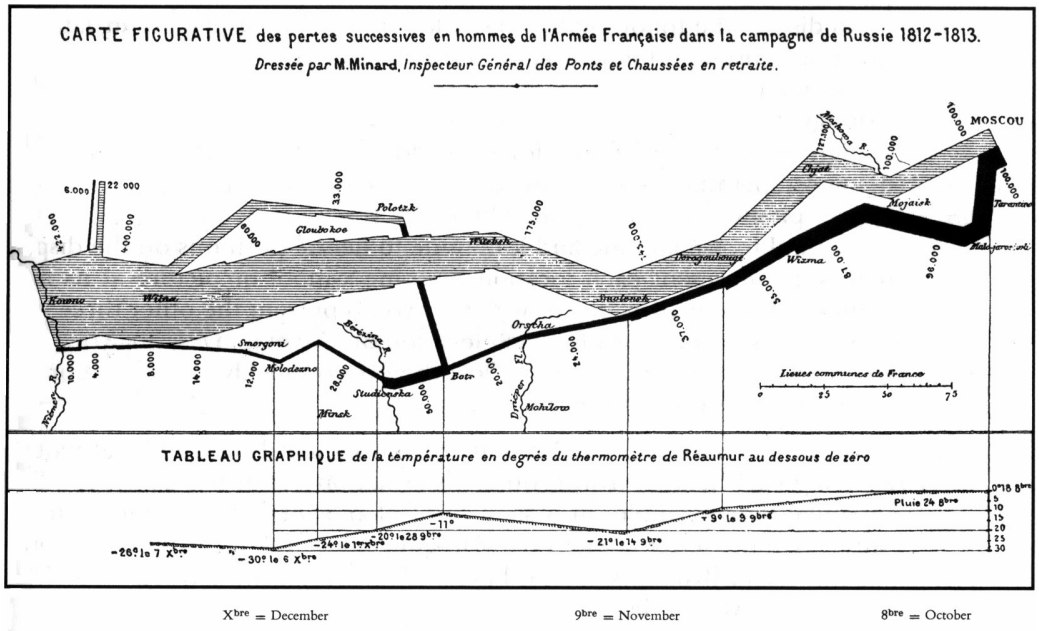


Abb. 1: Visualisierung erlaubt unserer Wahrnehmung schnell Zusammenhänge zwischen unterschiedlichen Parametern zu entnehmen (aus Wiley 1998).

VisuLab – Hilfsmittel für Daten-Detektive

Mit der Software *VisuLab* lernen Sie Techniken zur **Visualisierung mehrdimensionaler Daten** kennen, die über die Möglichkeiten von *Excel* hinausgehen. Es enthält zur Darstellung mehrdimensionaler Daten **vier Darstellungsarten**, die am besten nebeneinander verglichen werden. Es bietet somit eine Art Detektivhilfsmittel, um Daten grafisch nach Strukturen zu erkunden. *VisuLab* wird in der Regel aber nicht dazu eingesetzt, illustrative Grafiken zu erstellen. Dazu verwenden Sie auch in Zukunft mit Vorteil Excel oder andere Softwarepakete.

VisuLab ist aus einem Forschungsprojekt des Instituts für Computational Science der ETH Zürich hervorgegangen. Es ist frei erhältlich und kann auch auf Ihrem privaten Rechner installiert werden.

Vorgehen

Sie werden zuerst im *E.Tutorial* die einzelnen Darstellungsarten und einige Operationen von *VisuLab* kennen lernen, um damit in der Testaufgabe einen mehrdimensionalen Datensatz analysieren zu können.

Teil B: *E.Tutorial*

Arbeiten Sie das *E.Tutorial* Praxis 3 durch!

Sie finden das *E.Tutorial* auf Ihrer **CD-ROM** oder über **http://www.evim.ethz.ch**.

Im *E.Tutorial* Praxis 3 lernen Sie...

- ...Grafiken aus Excel-Tabellendaten erstellen (Lektionen 1 - 3)
- ...Daten von Excel nach VisuLab exportieren (Lektionen 4 + 5)
- ...4 verschiedene Grafiktypen in VisuLab erstellen und analysieren (Lektionen 6 - 9)
- ... Muster und Gesetzmässigkeiten in Daten erkennen mit Hilfe verschiedener VisuLab-Operationen, mit denen Sie Daten grafisch verarbeiten (Lektionen 10 - 13)

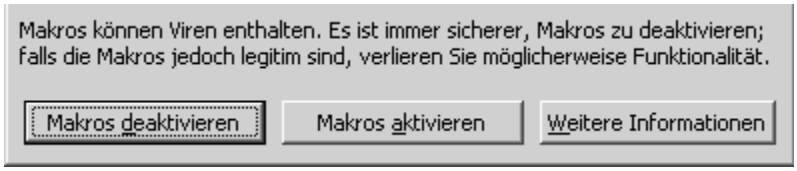
Benutzen Sie die **Erinnerungshilfe** (Reminder) auf den nächsten Seiten, um festzuhalten, wo Sie welche Funktion finden!

System-Voraussetzungen

PC mit MS Excel Version 97 oder höher. Für das Herunterladen der Beispieldateien brauchen Sie eine Internetverbindung.

Vorsicht: Makro-Sicherheit

Sie werden in diesem Praxismodul eine Datei, welche ein Makro enthält, starten müssen. Da sich viele Viren ebenfalls der Makrofunktion bedienen, wurden in die Office-Anwendungen verschiedene Sicherheitsüberwachungen integriert, so auch in Excel. Setzen Sie die **Makrosicherheit** für dieses Praxismodul unter Extras > Makro > Sicherheit auf **Mittel**, dann werden Sie vor dem Öffnen "makrohaltiger" Dokumente gefragt, ob Sie deren Ausführung zulassen wollen.



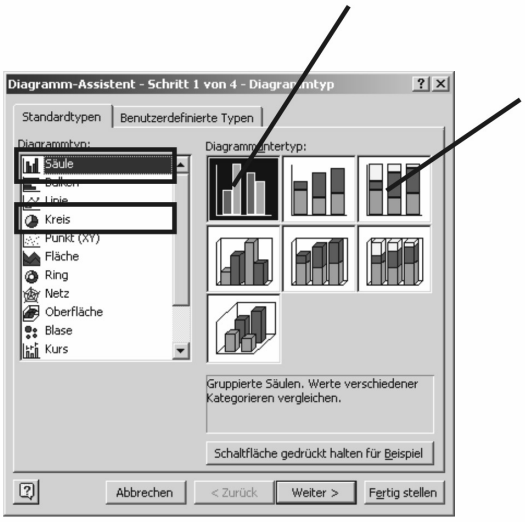
Dieses Fenster erscheint, wenn Sie die Makrosicherheit auf "Mittel" setzten. Somit können Sie wählen, ob Sie ein Makro aktivieren wollen oder nicht.

Weitere Hinweise finden Sie auf dem Blatt "Informationen zum Aufbau der Praxismodule"!

Reminder zum ***E.T****utorial* – wo finde ich was?

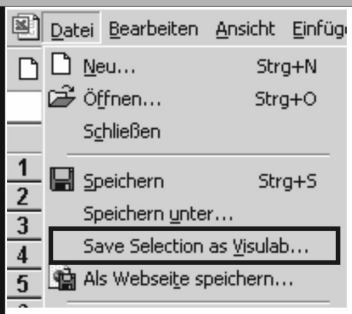
Visualisieren mit Excel

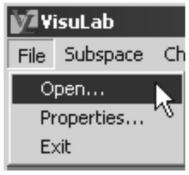
Zu Lektion 1 / 2

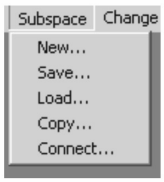


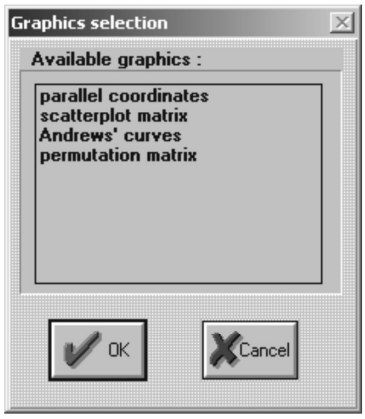
VisuLab öffnen/Dateien laden

Zu Lektion 5



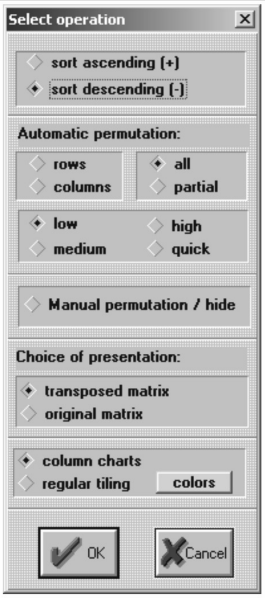


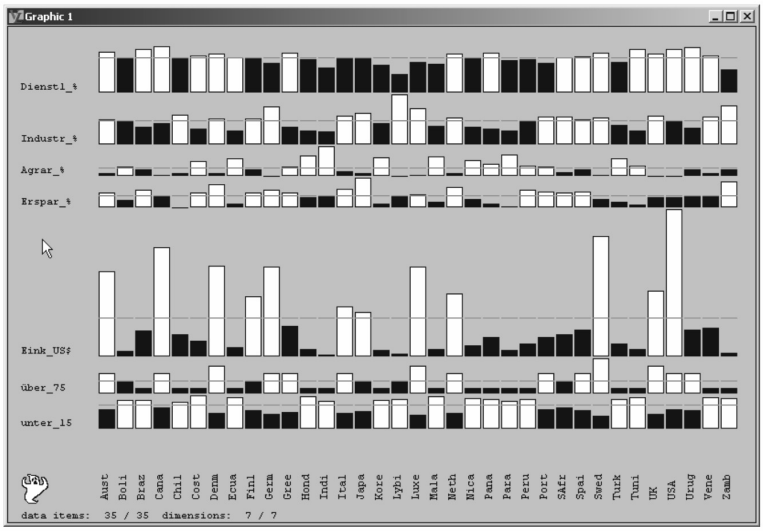




Permutationsmatrix

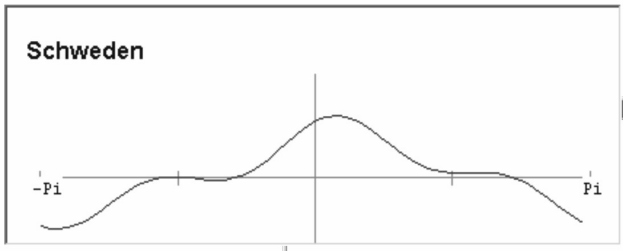
Zu Lektion 8

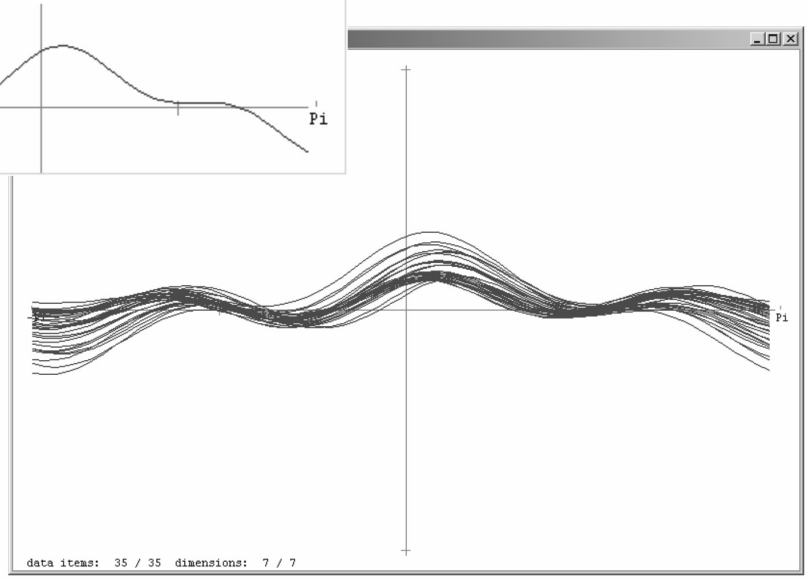




Andrews' Curves

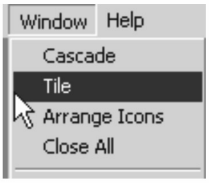
Zu Lektion 9





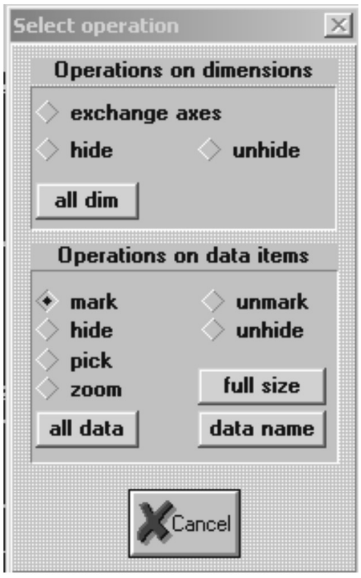
Explorative Datenanalyse mit VisuLab

Zu Lektion 10



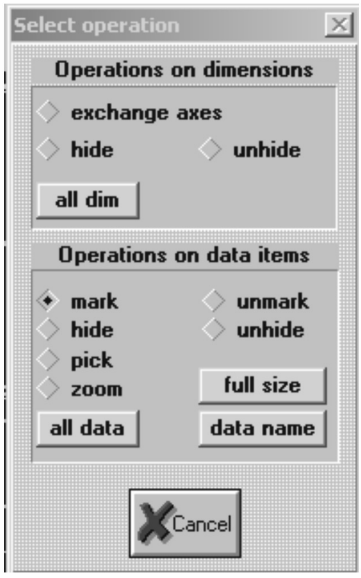
VisuLab-Operationen (1)

Zu Lektion 11



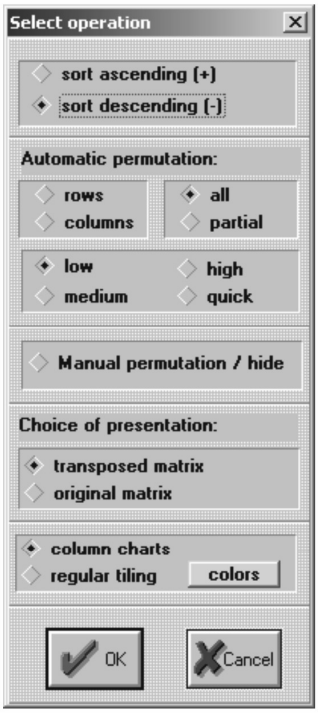
VisuLab-Operationen (2)

Zu Lektion 12



VisuLab-Operationen (3)

Zu Lektion 13



Teil C: Testaufgabe

1. Einführung

Sie haben im *E.Tutorial* gelernt, wie sie mehrdimensionale Daten mit *VisuLab* visualisieren können. In der Testaufgabe werden Sie mit dieser Software explorative Datenanalyse betreiben, um eine Hypothese aus der Biologie zu überprüfen.

Ein Gedankenmodell: Hybride und die Artenbildung in der Evolution

Hybride sind Nachkommen von Eltern *zweier verschiedener Arten* oder von *Eltern mit erblich verschiedenen Merkmalen*. Hybriden kommen natürlich vor und spielen für die genetische Vielfalt eine wichtige Rolle. Künstlich kann man sie erzeugen, indem die Geschlechtszellen verschiedenartiger Lebewesen kombiniert werden. Je näher die Eltern miteinander verwandt sind, desto besser gelingt die Herstellung einer Hybride. Unterscheiden sich die Eltern nur in den Merkmalsausprägungen eines oder mehrerer Gene (z.B. Fell- oder Blütenfarbe), entsteht meist eine lebensfähige, fruchtbare Hybride. Oft sind sie aber unfruchtbar. Es gibt Hybriden unterschiedlicher Pflanzenarten, die nach einer Verdoppelung des Chromosomensatzes (durch natürlich Mutation oder künstlichen Eingriff) eine neue fruchtbare Art bilden.

Bei der *Entstehung von neuen Arten* in der Evolution spielt die Hybridbildung ebenfalls eine Rolle. Isolationsmechanismen, wie z. B. Klimaveränderungen in der Vergangenheit (Eiszeiten), sorgen dafür, dass Kreuzungen zwischen verschiedenen Arten nicht möglich sind. Es folgt eine unabhängige genetische Entwicklung der getrennten Populationen bei unterschiedlichen Umwelt- und Selektionsbedingungen. Beide Genpools verändern sich unabhängig voneinander, so dass sich die getrennten Populationen zu unterschiedlichen Arten entwickeln. Bei der Aufhebung der Isolation (nach der Eiszeit beispielsweise) können sich die beiden Arten im Artenbildungsprozess noch so nahe stehen, dass eine Hybridenbildung möglich ist. Durch eine spontane Verdoppelung des Chromosomensatzes kann es auch vorkommen, dass eine fruchtbare Hybride entsteht, die sich im Lauf der Evolution als eine neue Art durchsetzen kann.

Die Untersuchungsobjekte: drei Schwertlilienarten *(Iris)*

- *Iris setosa*
- *Iris versicolor*
- *Iris virginica*

Links zu Bildern dieser Irisarten finden Sie auf unserer Homepage.

Hypothese und bisherige Befunde

Folgende Hypothese wurde 1934 von Randolph aufgestellt:

Die heutige Art *Iris versicolor* ist aus einer Hybride der beiden Arten *Iris setosa* und *Iris virginica* entstanden.

Folgende Befunde stützen bisher die Hypothese:

- Genetisches
- Verbreitung der Arten in Nordamerika
- Variation in Farbe und Form

a) Genetisches

Folgende Tabelle gibt die Zahl der Chromosomen und die Anzahl Chromosomensätze der drei Iris-Arten an. Der einfache (haploide) Chromosomensatz berechnet sich bei *Iris setosa* z.B. durch 38 / 2 = 19 Chromosomen.

	Anzahl Chromosomen	Anzahl Chromosomensätze
Iris setosa	38	2
Iris virginica	70	2
Iris versicolor	108	2

Bei der Entstehung der Art *Iris versicolor* aus einer Hybride der beiden anderen Arten, stellt man sich eine *Fusion* der beiden einfachen (haploiden) Chromosomensätze vor: 19+35 = 54 Chromosomen. Dieser "neue" Chromosomensatz kommt 2 mal vor, was 108 Chromosomen bei der heutigen Art *Iris versicolor* ergibt.

b) Verbreitung in Nordamerika

Die Abbildungen 2 bis 4 zeigen die Verbreitung der drei Irisarten *setosa, virginica* und *versicolor* in Nordamerika, die Abbildung 5 die Eisbedeckung des Kontinents während der letzten Eiszeit vor ca. 15'000 Jahren:

Abb. 2: Verbreitung von *Iris setosa* in Nordamerika

Abb. 3: Verbreitung von *Iris virginica* in Nordamerika

Abb. 4: Verbreitung von *Iris versicolor* in Nordamerika.

Abb. 5: Eisbedeckung von Nordamerika vor ca. 15'000 Jahren.

Iris setosa kommt heute im Nordwesten des Kontinents vor (*Abb. 2*), das in der letzten Einszeit wegen einer warmen Meeresströmung im Nordpazifik nicht vollständig eisbedeckt war (*Abb. 5*). Diese Verbreitung ist typisch für eine Pflanzenpopulation, die durch die Eismassen beinahe ausgerottet

wurde und in dieser Ecke des Kontinents überleben konnte. Durch einen freien Eiskorridor sollen übrigens auch die ersten Menschen von Sibirien her den amerikanischen Kontinent besiedelt haben.
Das Verbreitungsgebiet von *Iris virginica* ist im Osten und Südosten des Landes (*Abb. 3*). Während der Eiszeit konnte diese Pflanzenpopulation wohl nur im Südosten überleben und war somit vermutlich von der Population im Nordwesten isoliert worden. *Iris veriscolor* kommt heute ausschliesslich in Gebieten vor, die völlig eisbedeckt waren (*Abb. 4 und 5*).

c) Variation in Farbe und Form

Die Evolutionstheorie besagt, dass die Zunahme von Merkmalsunterschieden (z.B. Farbe oder Form) zwischen Individuen einer Art mit der Zeit zur Bildung von zwei Arten führen kann. Bei einer Hybriden-Art "mischen" sich die Unterschiede der beiden Elternarten zu einer *grösseren* Variabilität. Betrachtet man beispielsweise die *Farbenvariabilität* der drei Iris-Arten, stellt man bei *Iris versicolor* die grösste (lila-blau bis grün-gelb) und bei *Iris setosa* die kleinste (blau-grün) Variabilität fest.

Ein Datensatz, der die Variabilität dieser drei Arten auf der Basis ihrer Form zeigen kann, werden Sie nun in der folgenden Aufgabe untersuchen.

2. Vorbereitendes zur Testaufgabe

Datensammlung

Die beiden folgenden Excel-Dateien sind auf der Homepage zu finden:

Datei 1: ***Iris.xls*** enthält Daten von insgesamt **90 Irisblüten** aus einer Untersuchung an den drei Irisarten *Iris setosa, Iris versicolor und Iris virginica*. Die Datensammlung enthält Angaben über:

- die genaue Iris-Art (*Iris setosa, versicolor* oder *virginica*)
- die Länge und Breite der Kelchblätter (sepal-length und -width) in mm
- die Länge und Breite der Blütenblätter (petal-length und -width) in mm

Mit diesen Daten werden Sie *Unterscheidungskriterien zwischen den Arten* bestimmen.

Datei 2: ***Iris_unknown.xls*** enthält die gleichen Daten für **60 weitere Blüten**, dabei fehlt aber die Angabe der Iris-Art. Damit testen Sie, ob die mit Datei 1 aufgestellten Unterscheidungskriterien ausreichend präzise sind, um die Blütenart anhand ihrer Kelch- und Blütenblattdaten bestimmen zu können.

Die Abkürzungen nochmals im Überblick:

i-set	*Iris setosa*	sepal-leng	Kelchblatt-Länge
i-ver	*Iris versicolor*	sepal-widt	Kelchblatt-Breite
i-vir	*Iris virginica*	petal-leng	Kronblatt-Länge
		petal-widt	Kronblatt-Breite

So beginnen Sie die Aufgabe am besten...

- Laden Sie die beiden Excel-Dateien von der Homepage auf Ihren Rechner.
- Transformieren Sie die Excel-Datei ***Iris.xls*** in eine *VisuLab* Datei (z.B. ***Iris.vlb***).
- Stellen Sie die Daten in ***allen vier Darstellungsarten*** von *VisuLab* dar (verbunden mit **Subspace > Connect...**).
- Lassen Sie alle 4 Darstellungstypen gleichzeitig auf dem Bildschirm anzeigen (durch **Window > Tile**).

3. Aufgaben

Mit den gemessenen 4 Dimensionen haben Sie ein Mass für die Variabilität der Irisarten zur Verfügung, um damit möglicherweise Hinweise auf deren Verwandtschaft zu erhalten. 2 Teilaufgaben sind zu lösen:

Teil A: Kriterien bestimmen: Mit Hilfe von *VisuLab* suchen Sie unter den gemessenen Parametern mögliche Kriterien, wie Sie die Irisarten auseinanderhalten können.

Teil B: Kriterien testen: Sie testen die im Teil A aufgestellten Kriterien anhand von Iris-Proben unbekannter Art.

Teil A: Kriterien bestimmen (iris.xls)

- Suchen Sie mit Hilfe der visuellen Datenanalyse nach geeigneten **Dimensionen**, welche als Kriterien zur Bestimmung der 3 Iris-Arten eingesetzt werden können (**Abb. 6**).
- Bestimmen Sie **2 Kriterien** (inkl. Wert), die ein möglichst genaues Auseinanderhalten dieser drei Irisarten ermöglichen (**Abb. 6**). Bewerten Sie zusätzlich die Qualität der Kriterien. Kriterien können die Verteilung der beiden Arten vollständig teilen oder es entstehen überlappende Bereiche (siehe **Abb. 7**).

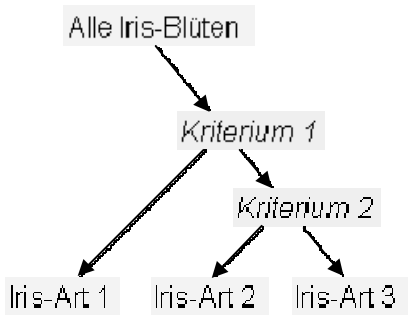


Abb. 6: Das Festlegen von zwei Kriterien in einer der gemessenen Dimensionen ermöglicht das Bestimmen der 3 Irisarten.

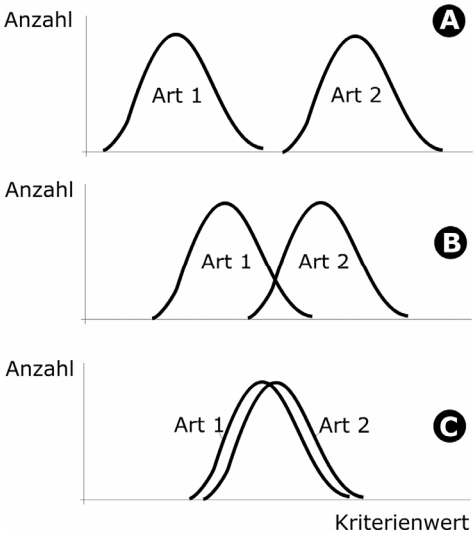


Abb. 7: Verteilung von zwei Irisarten entlang eines Kriterienwerts. Bei A können alle Arten eindeutig bestimmt werden, bei B und C gibt es einen überlappenden Bereich, wo die Arten nicht auseinander gehalten werden können.

Teil B: Kriterien testen (iris_unknown.xls)

- Bestimmen Sie mit den von Ihnen unter A aufgestellten Kriterien im Entscheidungsbaum die Irisarten der Datensätze der Datei ***Iris_unknown.xls***. Sie enthält je 20 Blüten einer Art.
- Lassen Sie sich die Nummern der Iris-Blüten der Datei *Iris_unknown.xls* auf dem Bildschirm anzeigen.

4. Form und Bedingungen

- Führen Sie einer Assistentin oder einem Assistenten mit *VisuLab* vor, wie Sie die beiden Kriterien gefunden haben. Begründen Sie auch die Qualität der Kriterien (siehe Abbildung 7). Wie können Sie trotz überlappender Bereiche genaue Kriterien zur Unterscheidung der Arten angeben?
- Begründen Sie, welche Darstellungsart sich für welchen Schritt besonders eignet und erklären Sie, was Sie daraus lesen können.
- Die Begriffe dieses Praxismoduls sollten Sie mit einfachen Worten erklären können.

5. FAQ's zur Testaufgabe

Können VisuLab-Darstellungen gespeichert werden?

Nein. Die Darstellungen müssen aus der geladenen *.vlb-Datei erstellt werden.

6. Literatur

Dutch, S. *Pleistocene Glaciers and Geography*.
http://www.uwgb.edu/dutchs/EarthSC202Notes/GLACgeog.htm
Fischer, R.A. *The use of multiple measurements in taxonomic problems*. 1. Iris Data. Ann. Eugenics 7 (1936).
Wiley, J. *Understanding Data*. Jacaranda Wiley (1998).