Mitteilungen Klosterneuburg 57 (2007): 3-22

Herkunftsunterscheidung von Weinen der Sorte 'Grüner Veltliner' anhand ihrer Aromaprofile mit Hilfe von Data Mining-Techniken und Neuronalen Netzwerken

Andreas Schreiner[1], Walter Brandes[1], Giuseppe Versini[2], Emmerich Berghofer[3] und Reinhard Eder[1]

[1]Höhere Bundeslehranstalt und Bundesamt für Wein- und Obstbau
A-3400 Klosterneuburg Wiener Straße 74

[2]Istituto Agrario San Michele all'Adige
I-38010 San Michele all'Adige (TN), via Edmondo Mach, 1

[3]Universität für Bodenkultur
Department für Lebensmittelwissenschaften und -technologie
A-1190 Wien, Muthgasse 18

Im Rahmen vorliegender Arbeit wurde getestet, inwieweit anhand von Aromaanalysen bei österreichischen Weinen der Sorte 'Grüner Veltliner' eine korrekte Zuordnung der Herkunft (Weinbaugebiet) möglich ist. Die Aromaanalyse bestand aus einer Festphasenextraktion zur Trennung und Konzentrierung der freien und glykosidisch gebundenen Komponenten und anschließender Bestimmung mittels Gaschromatographie und Flammenionisationsdetektor. Die Konzentrationen von 176 Komponenten konnten bestimmt und die Aromasubstanzen zum Teil identifiziert werden. Zur Herkunftserkennung wurden 59 Weine der Sorte 'Grüner Veltliner', Jahrgang 2004, aus der weinbaulichen Praxis analysiert. Die Ergebnisse wurden mit Hilfe von statistischen und Data Mining- Techniken untersucht. Die Überprüfung der Tauglichkeit der einzelnen Methoden wurde über eine probeweise Vorhersage aller Proben mit Hilfe der leave one out-Methode durchgeführt. An statistischen Verfahren wurde der k-nächste Nachbarn (knn) Vergleich, die Lineare Diskriminanzanalyse (LDA) und ein Neuronales Netzwerk eingesetzt. Je nach Methode und Vorbereitung des Datensatzes konnten Trefferquoten zwischen 40,7 und 81,8% erzielt werden. Dabei zeigte sich, dass Weine aus Weinbaugebieten, von denen nur eine geringe Probenanzahl verfügbar war, deutlich schlechter vorhergesagt wurden. Die am besten geeignete Vorhersagemethode war das Neuronale Netzwerke mit einer Trefferquote von über 80%. Für den praktischen Einsatz der Methode ist es jedoch erforderlich, genügend Proben aus den zur Auswahl stehenden Gebieten zu analysieren, um damit Vergleichswerte und Trainingsdaten für die Zuordnung unbekannter Proben zu erstellen.

Schlagwörter: Wein, 'Grüner Veltliner', Aromaanalyse, Herkunftsbestimmung, Data Mining, Lineare Diskriminanzanalyse, Neuronale Netzwerke

Differentiation of origins of 'Grüner Veltliner' wines by means of aroma profiles and data mining techniques and neural networks. *In this study, aroma analysis was carried out with Austrian wines of the cultivar 'Grüner Veltliner' with the intention to predict their origin (i.e. the specific growing area in Austria). The principle of the analysis was a solid phase extraction of the volatile compounds of a given wine sample, followed by separation and concentration of the extracted free and glycosidically bound components by means of GC and FID. The concentrations of a total of 176 components were measured and their aroma substances were partially identified. The results were examined by statistical and data mining techniques. The leave one out-technique was used to check the efficiency of the particular methods. The statistical methods used were the k-Next Neighbour comparison (knn), the Linear Discriminant Analysis (LDA) and a Neural Network. Depending on the method and the preparation of the datasets, hit ratios*

ranging from 40.7 to 81.8% were reached. It was shown that a correct prediction of origin was much more unlikely for wines originating from a growing region from which only a few samples have been analysed. The application of a Neural Network emerged as the most suitable of all examined prediction methods (hit ratios up to about 80%). For a possible deployment of the method into day to day analytical practice it would be advisable to ensure the availability of an adequate number of samples from all winegrowing regions in question. This is necessary to provide comparison and training data for the prediction of unknown samples.
Key words: Wine, 'Grüner Veltliner', aroma analysis, determination of origin, Data Mining, linear discriminative analysis, neural networks

La distinction des origines de vins du cépage 'Grüner Veltliner' sur la base de leurs profils aromatiques à l'aide des techniques d'extraction des données (data mining) et des réseaux neuronaux. *Dans le cadre du présent travail, il a été testé dans quelle mesure une attribution correcte de l'origine (région viticole) au moyen d'analyses des arômes est possible pour les vins autrichiens du cépage 'Grüner Veltliner'. L'analyse des arômes consistait en une extraction en phase solide destinée à la séparation et à la concentration des composants libres et de ceux faisant l'objet d'une liaison glycosidique, suivi de la détermination à l'aide de la chromatographie en phase gazeuse et du détecteur à ionisation de flamme. Il a été possible de déterminer les concentrations de 176 composants et d'identifier partiellement les substances aromatiques. 59 vins du cépage 'Grüner Veltliner', millésime 2004, issus de la pratique vinicole, ont été analysés. Les résultats ont été examinés par des moyens statistiques et des techniques d'extraction des données (data mining). La vérification de l'aptitude des différentes méthodes a été effectuée en réalisant des pronostics de l'ensemble des échantillons à titre d'essai à l'aide de la méthode leave one out. Les procédures statistiques utilisées ont été la comparaison des k plus proches voisins (kppv), l'analyse discriminante linéaire (ADL) et un réseau neuronal. Des taux de pertinence entre 40,7 et 81,8 % ont pu être obtenus en fonction de la méthode et de la préparation des enregistrements. Dans ce contexte, il s'est révélé que les pronostics pour des vins en provenance de régions viticoles, pour lesquelles on ne disposait que d'un faible nombre d'échantillons, avaient été beaucoup moins corrects. La méthode de pronostic la plus appropriée avec un taux de pertinence de plus de 80 % a été celle des réseaux neuronaux. Pour la mise en pratique de cette méthode, il est cependant nécessaire d'analyser une quantité suffisante d'échantillons provenant des régions à sélectionner, afin d'établir des valeurs de comparaison et des données d'apprentissage pour le classement d'échantillons inconnus.*
Mots clés : vin, 'Grüner Veltliner', analyse des arômes, détermination de l'origine, data mining, analyse discriminante linéaire, réseaux neuronaux

Seit langem werden Möglichkeiten erforscht, Weine mit Hilfe von analytischen und sensorischen Parametern zu klassifizieren. Häufig verwendete Unterscheidungskriterien sind beispielsweise die Qualitätstufe, die Herkunft eines Weines oder die Rebsorte. Im Bereich der Weinaromaforschung gibt es dabei langjährige Erfahrungen. So gelang bereits im Jahr 1978 RAPP und HASTRICH (1978) die Unterscheidung der Rebsorten 'Riesling' und 'Morio-Muskat' mit Hilfe einer multiplen Diskriminanzanalyse von Messwerten der Gehalte von 27 ausgewählten Aromakomponenten. Sie erkannten außerdem, dass die Herkunft einen, wenn auch untergeordneten, Einfluss auf die Aromakomposition von Weinen derselben Sorte ausübt. In späteren Untersuchungen gelang eine Differenzierung von verschiedenen Weinen, die zwar alle die Bezeichnung 'Riesling' tragen, aber nicht mit dem 'Weißen Riesling' verwandt sind. Für diese Differenzierung waren nur zwölf sortenspezifische Terpene ausreichend (RAPP et al., 1985).

Andere Arbeiten zielten darauf ab, Weine einzelnen Jahrgängen zuzuordnen. So konnten SEEBER et al. (1991) Moste und Weine der Sorte 'Chardonnay' anhand der Konzentrationen von Aminosäuren (in Most und Wein), Metallionen (Wein) und von flüchtigen Inhaltsstoffen (Wein) erfolgreich drei unterschiedlichen Jahrgängen zuordnen. ARRHENIUS et al. (1996) untersuchten die regionalen Unterschiede von Chardonnay-Weinen aus verschiedenen kalifornischen Anbaugebieten. Dabei kamen sowohl Methoden der beschreibenden Sensorik als auch eine analytische Bestimmung der Aromastoffgehalte mittels GC-MS zum Einsatz. In weiteren Untersuchungen zur Herkunftsbestimmung von Weinen mit Hilfe von analytischen Parametern betrachtete BERENTE (2004) die Anthocyangehalte deutscher Rotweine. KLIMMEK (2003) verglich in einer sehr umfangreichen Arbeit die Eignung zahlreicher Parameter, wie zum Beispiel Stabilisotopengehalte, Mineralstoffgehalte, flüchtige Inhaltsstoffe und Anthocyanzu-

sammensetzung, zur Unterscheidung der Herkunft von einigen Hundert Weinen aus der ganzen Welt.
Das Ziel der meisten dieser Untersuchungen ist es, geeignete Mittel für die Qualitätssicherung und Authentizitätsüberprüfung zu finden. So sollen Angaben zu Weinen überprüfbar werden. Neue Methoden, die dies ermöglichen, werden vor allem dort gesucht, wo man sich bisher auf das natürlicherweise subjektive Urteil von Prüfpersonen im Zuge sensorischer Untersuchungen stützen musste. Auch als Ergänzung zu bereits etablierten, aber sehr aufwändigen und teuren Analyseverfahren, wie der Herkunftserkennung über Stabilisotopenanalyse, sind Alternativen hoch interessant (Klimmek, 2003; Berente, 2004). Die aromaaktiven flüchtigen Inhaltsstoffe, die einen wichtigen Qualitätsfaktor des Weines darstellen, sind daher ein ideales Forschungsobjekt.

Material und Methoden

'Grüner Veltliner'

Der 'Grüne Veltliner' ist mit einer Rebfläche von knapp 17.500 ha und damit etwa 36% der Gesamtweinbaufläche die in Österreich am weitesten verbreitete Rebsorte (ÖWM; 2005). Trotzdem ist über die flüchtigen aromaaktiven Inhaltsstoffe dieser Rebsorte noch wenig bekannt. Die Analytik von Weinaromen beschäftigte sich bisher hauptsächlich mit Sorten, die entweder eine große internationale Verbreitung haben oder die einen besonders hohen Gehalt an sortentypischen primären Aromakomponenten aufweisen. Beispiele dafür sind natürlich der 'Weiße Riesling', gefolgt von Muskat-Sorten und den Traminer-Typen (Versini et al., 1981; Rapp et al., 1985). Der 'Grüne Veltliner' fristete in dieser Beziehung bisher ein Schattendasein, da die meisten dieser Untersuchungen nicht in Österreich durchgeführt wurden und der 'Grüne Veltliner' bis auf wenige Ausnahmen in benachbarten Ländern (Tschechische Republik, Ungarn) eine rein österreichische Spezialität ist.
Der 'Grüne Veltliner' ist international als typisch österreichische Sorte etabliert, die einzelnen Anbaugebiete sind allerdings auf Grund ihrer geringen Größe weitgehend unbekannt. Im österreichischen Markt wird hingegen viel Wert auf die genaue Herkunft der Weine gelegt. Produzenten, Vermarkter und ganze Anbauregionen versuchen durch Hervorhebung von einzelnen erhofften oder tatsächlich vorhandenen regionsspezifischen Eigenschaften ihrer Weine ein markantes, für den Konsumenten wieder erkennbares Profil zu gewinnen. Dies führt zu der Frage, inwieweit diese Differenzierungen auf tatsächlich beim Weingenuss bemerkbare Unterschiede zurückzuführen sind. Bisher konnte man dieser nur mit Hilfe von sensorischen Prüfungen, mit all den Nachteilen, die der Einsatz von menschlichen Prüfern mit sich bringt, nachgehen. Die vorliegende Arbeit soll nun einen Beitrag dazu leisten, diese Fragestellung durch den Einsatz aromaanalytischer Methoden zu klären.

Weinaroma

Das Weinaroma umfasst alle flüchtigen Inhaltsstoffe eines Weines, die vom sensorischen Organ des menschlichen Geruchssinnes wahrgenommen werden können (Clarke und Bakker, 2004; Rapp, 1998). Nach derzeitiger Einschätzung sind dies zumindest 600 bis 800 unterschiedliche Substanzen, die insgesamt in sehr geringen Konzentrationen das Bukett eines Weines ausmachen. Die Gesamtkonzentration aller Aromastoffe liegt bei etwa 0,8 bis 1,2 g/l, wobei die Konzentration einzelner Substanzen üblicherweise in einem Bereich von 10^{-2} bis 10^{-10} g/l liegt (Rapp und Mandary, 1986; Rapp, 1998). Die im Wein vorkommenden Aromastoffe wurden von Rapp (1998) in primäres und sekundäres Traubenaroma, Fermentationsaroma und Reifungsaroma eingeteilt. Die wichtigsten Komponenten des primären Traubenaromas sind Monoterpene, dazu gehören noch Norisoprenoide, Phenole und aliphatische bzw. methylverzweigte Verbindungen (Rapp, 1998; Papargyriou, 2003). Weitere Verbindungen, die zum Teil erst vor kurzem als wichtige Aromakomponenten in Wein erkannt wurden, sind Pyrazine und Mercaptane (Amann, 2003). Die sortentypischen Eigenschaften von Weinen werden zu einem großen Teil von ihrem Gehalt an primären Aromakomponenten bestimmt. Die in Trauben vorkommenden primären Aromakomponenten und ihre Konzentrationen sind stark von der Rebsorte abhängig. So konnten zum Beispiel Sorten mit hohen Terpengehalten mit großer Sicherheit über ihren Gehalt an verschiedenen Monoterpenen identifiziert und unterschieden werden (Rapp und Hastrich, 1978; Rapp et al., 1985).
Zum sekundären Traubenaroma zählen Komponenten, die bei der Traubenverarbeitung durch chemische, enzymatische oder thermische Reaktionen aus Traubeninhaltsstoffen gebildet werden. Ein Großteil der dabei auftretenden Veränderungen von Inhaltsstoffen ist auf die Hydrolyse von glykosidisch gebundenen Formen der primären Traubenaromakomponenten zurückzuführen (Williams et al., 1981; Rapp, 1988).

Aromakomponenten, die bei der Vergärung des Mostes gebildet werden, bestimmen das Fermentationsbukett und bilden zusammen den anteilsmäßig größten Teil der im Wein vorkommenden Aromakomponenten. Die meisten davon entstammen dem Stoffwechsel der zur Fermentation eingesetzten Hefe und machen bis zu 50% der Gesamtmenge der aromaaktiven Substanzen eines Weines aus (RAPP, 1998). Die Mengen und Anteile der gebildeten Substanzen, hauptsächlich höhere Alkohole (Fuselöle), Diole, Ester, Säuren, Aldehyde, Ketone und Schwefelverbindungen, hängen stark von den bei der Gärung herrschenden Bedingungen, wie Temperatur und Ernährungszustand der Hefe, vorhandenen Präkursoren und natürlich auch vom Hefestamm selbst ab (WONDRA und BEROVI, 2001; MAJDAK et al., 2002). In den letzten Jahren wurde bekannt, dass zusätzlich zur Bildung der genannten Substanzen im Zuge der Fermentation auch eine Umwandlung nennenswerter Anteile von primären Traubenaromen auftreten kann. So berichteten KING und DICKINSON (2003) von einer zum Teil beträchtlichen Umwandlung von Terpenen durch Hefe bei Gärversuchen in Bier.

Während der Reifung eines Weines in Fass, Tank oder Flasche treten weitere Reaktionen auf, die einen Einfluss auf die Aromakomposition ausüben. Die dabei entstehenden olfaktorischen Eindrücke werden als Reifungsaroma bezeichnet und durch Abbaureaktionen und Umwandlungen von im Wein enthaltenen Komponenten verursacht. Eine wichtige hier zu nennende Substanzgruppe sind Norisoprenoide, wie etwa β-Damascenon oder Vitispiran, die unter anderem einen entscheidenden Einfluss auf die Alterung von Riesling-Weinen ausüben (RAPP et al., 1985; WALLNER et al. 1999).

Aromaanalytik

Für die Analytik flüchtiger Inhaltsstoffe von Wein nimmt die Gaschromatographie auf Grund ihrer Fähigkeit zur Trennung und zum Nachweis von hunderten der in Frage kommenden Inhaltsstoffe einen entscheidenden Stellenwert ein. Bei der Probenaufbereitung dienen verschiedenste Methoden zur Abtrennung der zu analysierenden Substanzen aus dem Ursprungswein. Zur Detektion und/oder Identifikation der Aromakomponenten wird die Gaschromatographie mit einer großen Vielfalt an Detektionsmethoden, vom Flammenionisationsdetektor (FID) bis hin zur Massenspektroskopie (GC-MS) und Olfaktometrie kombiniert (FERREIRA et al., 2002; QUADT, 1999).

Bereits kurz nach den Anfängen der Aromaanalytik in den 60iger-Jahren wurden die Terpene als die für die Sortencharakteristik einiger aromatischer Weinsorten ausschlaggebende Gruppe von Weininhaltsstoffen erkannt und ausgiebig erforscht. So beschrieben RAPP und HASTRICH (1978) sowie VERSINI et al. (1981) die Terpenzusammensetzung von Weinen der Sorte 'Rheinriesling', DI STEFANO (1981) untersuchte die Sorte 'Gewürztraminer'. In den folgenden Jahren verlagerte sich der Schwerpunkt der Forschungen in spezifischere Themenbereiche. So wurden die im Wein vorkommenden Terpene genau charakterisiert und dabei auch immer wieder neue, zuvor unbekannte Verbindungen entdeckt sowie weitere Rebsorten charakterisiert. Ein weiterer wichtiger Forschungsschwerpunkt wurde die Klassifizierung von Weinen auf Grund ihrer Aromastoffzusammensetzung. Ermöglicht wurde dies durch die fortlaufende Weiterentwicklung der eingesetzten Analysenmethoden und der technischen Ausstattung. Während zu Beginn aufwändige flüssig-flüssig Probenextraktionen mit zum Teil sehr problematischen Lösungsmitteln, wie Freon, vorgenommen wurden, setzten sich in den letzten 15 Jahren immer mehr Techniken der Festphasenextraktion (Solid Phase Extraction, SPE) in verschiedenen Varianten (Solid Phase Micro Extraction, Stir Bar Extraction) zur Anreicherung der zu untersuchenden Komponenten durch. Diese zeichnen sich durch einen weitaus geringeren Lösungsmittelverbrauch aus, sind allgemein schneller und einfacher durchzuführen und ermöglichen zum Teil sogar eine Analyse der flüchtigen Weininhaltsstoffe ohne vorhergehende Probenvorbereitung. Auch im Bereich der Detektionsmethoden wurden entscheidende Fortschritte erzielt. Während der klassische Detektor für Aromastoffe, der sehr unspezifische Flammenionisationsdetektor, auf Grund seiner Empfindlichkeit und der Eignung zur Konzentrationsbestimmung über interne Standards immer noch häufig im Einsatz ist, wird heute vermehrt eine GC-MS-Kopplung eingesetzt. Nach der Erstellung spezifischer Fragmentbibliotheken mit Hilfe von Standards ist die GC-MS-Kopplung ein höchst empfindliches Detektionsverfahren zur Aromaanalyse, das es in manchen Fällen sogar ermöglicht, zwei chromatographisch nicht trennbare Substanzen zu identifizieren, in diesem Fall allerdings nicht zu quantifizieren. Viele dieser Methoden ermöglichen zumindest eine eingeschränkte Quantifizierung der messbaren Aromakomponenten und öffnen damit ein weites Feld für Anwendungen der Aromaanalytik. Zusätzliche Informationen, vor allem in Bezug auf den sensorischen Einfluss einzelner Substan-

zen, lassen sich durch einen olfaktorischen Detektor erhalten.

Statistik und Data Mining

Um die durch Aromaanalysen gewonnenen Daten zur Klassifizierung der Weine nach gewünschten Kriterien verwenden zu können, werden üblicherweise statistische Methoden eingesetzt, die es ermöglichen, verwertbare Ergebnisse aus den anfallenden großen Datenmengen zu erhalten. Bei sehr großen Datenmengen benötigt man dazu oft die Hilfe von Data Mining-Techniken. Diese Auswertungsstrategie, die ihrerseits wieder auf statistische Methoden, oft noch kombiniert mit Techniken des Machine Learning und diversen Mustererkennungsstrategien, zurückgreift, ermöglicht eine Extraktion relevanter Daten und Zusammenhänge aus einer unübersichtlich großen Menge von Rohdaten. Das erlaubt die Durchforstung von beinahe beliebig großen Datensätzen nach enthaltenen Informationen und Mustern beziehungsweise Abhängigkeiten zwischen verschiedenen Faktoren (www.liacc.up.pt/~ltorgo/DataMiningWithR/; www.statsoft.com/textbook/stathome.html,). Die Vorgangsweise beim Data Mining lässt sich in drei Schritte einteilen, die im Folgenden kurz dargestellt werden.

Explorative Datenanalyse. In der explorativen Datenanalyse werden die Daten für die weitere Bearbeitung aufbereitet. Darunter fällt zum Beispiel die Entfernung von fehlerhaften oder aus einem anderen Grund problematischen Datensätzen beziehungsweise die Ergänzung von fehlenden Daten. Bei sehr großen Datenmengen ist es außerdem oft sinnvoll, einen Teil der Datensätze auszuwählen, um eine handhabbare Menge für die nachfolgend angewendeten statistischen Verfahren zu erhalten. Dabei können verschiedene visuelle und statistische Methoden zur Erkennung der relevanten Daten und Variablen verwendet werden (www.liacc.up.pt/~ltorgo/DataMiningWithR/; www.statsoft.com/textbook/stathome.html).

Erstellung und Validierung eines Modells. Auf die explorative Datenanalyse folgt die Erstellung und Validierung eines oder mehrerer Modelle zur Beschreibung der Messdaten. Diese Modelle werden dann auf ihre Qualität überprüft. Als Kriterien dienen dabei zum Beispiel die Vorhersagegenauigkeit sowie die Reproduzierbarkeit und Stabilität bei wiederholten Vorhersagen verschiedener Testdatensätze. Dieser oftmals sehr aufwändige Vorgang involviert normalerweise den Vergleich mehrerer erstellter Modelle mit Hilfe von statistischen Vergleichsverfahren (www.liacc.up.pt/~ltorgo/DataMiningWithR/; www.statsoft.com/textbook/stathome.html).

Anwendung des Modells. Nachdem das leistungsfähigste Modell erkannt wurde, kann es in der Folge auf neue Daten angewendet werden, um möglichst genaue Vorhersagen zu treffen (www.liacc.up.pt/~ltorgo/DataMiningWithR/; www.statsoft.com/textbook/stathome.html).

Das Konzept des Data Mining verbreitet sich zunehmend in verschiedensten Forschungsgebieten. Am stärksten wachsen dabei Anwendungen im wirtschaftlichen und im Finanzbereich (Zirilli, 1997). Dadurch wurden in den letzten Jahren viele speziell darauf zugeschnittene Analysetechniken im Bereich des Data Mining entwickelt. Trotzdem sind traditionellere statistische Verfahren, wie die Explorative Datenanalyse und Modellierungstechniken weit verbreitet. Gerade die Explorative Datenanalyse ist, wie oben gezeigt wurde, ein wichtiger, oft auch eigenständig eingesetzter Teilbereich des Data Mining. Letzteres ist jedoch viel weiter gefasst und legt vor allem viel Wert auf praktische Anwendung der gewonnenen Erkenntnisse, dafür aber weniger auf die genaue Erklärung der zugrunde liegenden Zusammenhänge und Gesetzmäßigkeiten. Auf Grund dieser Zielsetzung sind im Bereich des Data Mining auch Black Box-Systeme, bei denen man nicht genau über die bei der Erkenntnisgewinnung ablaufenden Vorgänge Bescheid weiß, seit längerem akzeptiert, während sie diese Akzeptanz in anderen Bereichen der Statistik nicht genießen. Ein Beispiel dafür sind die weiter unten beschriebenen Neuronalen Netzwerke. Auf Grund ihrer robusten und zuverlässigen Ergebnisse finden diese Techniken aber auch im naturwissenschaftlichen Bereich mehr und mehr Anwendungen (www.liacc.up.pt/~ltorgo/DataMiningWithR/; www.statsoft.com/textbook/stathome.html).

Im Verlauf der vorliegenden Arbeit wurden einige statistische Methoden zur Auswertung und für das Data Mining herangezogen, die hier ein wenig genauer erläutert werden sollen. Zum Einsatz kamen die einfaktorielle Varianzanalyse (Quadt, 1999), die Hauptkomponentenanalyse (Henrion und Henrion, 1995; Berente, 2004), der k-nächste Nachbarn Vergleich sowie die Lineare Diskriminanzanalyse nach Anleitungen von Klimmek (2003) und Henrion und Henrion (2004) und ein Neuronales Netzwerk wie im Electronic Statistics Textbook der R-Foundation (www.statsoft.com/textbook/stathome.html) beschrieben.

Varianzanalyse. Bei der einfaktoriellen Varianzanalyse werden die Streuungen um den Mittelwert einer

Untergruppe der Probengesamtheit mit den Unterschieden zwischen Mittelwerten aller Gruppen verglichen. Dieser Vergleich der Inner-Gruppen-Varianzen mit der Zwischen-Gruppen-Varianz ermöglicht eine Aussage darüber, ob und mit welcher Signifikanz sich die Gruppen voneinander unterscheiden. Signifikante Unterschiede liegen dann vor, wenn die Streuung um die Gruppenmittelwerte klein im Vergleich zur Streuung zwischen den Mittelwerten ausfällt. Obwohl der Einsatz der Varianzanalyse nach einer Reihe von vorgegebenen Bedingungen verlangt, die im praktischen Einsatz vielfach nicht voll gewährleistet werden können, hat sie sich trotzdem als robuste Methode erwiesen und wird verbreitet eingesetzt (QUADT, 1999).

Hauptkomponentenanalyse (PCA). Die Hauptkomponentenanalyse ist im Gegensatz zur Varianzanalyse ein unüberwachtes statistisches Verfahren. Unüberwachte Verfahren sind solche, die Unterschiede bzw. Strukturen im Datensatz ausschließlich auf Grund von Variablenwerten aufzeigen (HENRION und HENRION, 1995; BERENTE, 2004). Das Ziel der PCA ist, die in einem Datensatz vorhandenen Varianzen in wenigen neu errechneten Variablen möglichst vollständig auszudrükken. Dazu werden viele Variablen (Komponenten) zu übergeordneten Faktoren (Hauptkomponenten) zusammengefasst, die einen großen Teil der ursprünglich vorhandenen Varianz und damit der enthaltenden Information aufnehmen (HENRION und HENRION, 1995; KLIMMEK, 2003; BERENTE, 2004). Da diese neu kombinierten Faktoren selten die gesamte ursprüngliche Varianz aufnehmen können, geht ein Teil der vorhandenen Information verloren. Das Ziel der PCA ist nun, die Faktoren so anzuordnen, dass dieser Verlust minimiert wird. Die optimale Anordnung erreicht man durch Rotation der Achsen des Koordinatensystems, in dem die Faktoren abgebildet werden, so lange, bis die enthaltene Varianz ein Maximum erreicht.

knn-Vergleich. Die Methode der *k*-nächsten Nachbarn (knn) ist eine der einfachsten Klassifikationsmethoden und gehört zu den überwachten Verfahren. Das bedeutet, dass im Gegensatz zu den unüberwachten Verfahren keine Klassifizierung ausschließlich auf Grund der bekannten Beobachtungswerte möglich ist. Überwachte Verfahren müssen mit einer repräsentativen Auswahl an Proben trainiert werden, mit deren Hilfe dann ein Klassifikationsmodell erstellt wird. Nachfolgend wird dieses Klassifikationsmodell umfangreichen Tests mit bekannten Proben unterzogen und erst danach ist eine Klassifizierung unbekannter Proben Erfolg versprechend. Die Überprüfung des Modells erfolgt dabei häufig nach der leave one out-Methode. Bei dieser wird jeweils eine einzelne Probe eines Probenpools ausgelassen, während das Klassifizierungsmodell mit Hilfe der übrigen Proben erstellt wird. Die verbliebene Probe wird danach anhand des Modells der Klasse zugeteilt, mit der sie die meisten Gemeinsamkeiten aufweist, und die Korrektheit dieser Zuteilung wird überprüft. Durch wiederholte Anwendung dieser Methode auf alle Proben des Pools kann die Qualität des Klassifizierungsmodells überprüft werden. Diese Überprüfung dient außerdem zur Vermeidung des overfitting-Effekts, der bei einer zu starken Anpassung des Klassifikationsmodells an den zum Training verwendeten Probenpool auftritt. Das Modell stützt sich in diesem Fall auf zufällige im Trainingspool enthaltene Merkmale, die es ermöglichen, alle darin enthaltenen Proben korrekt zu klassifizieren, die aber nicht auf andere Proben außerhalb des Trainingspools übertragbar sind. Die leave one out-Methode ist eine gute Möglichkeit, das overfitting zu erkennen, um gegebenenfalls Gegenmaßnahmen ergreifen zu können (KLIMMEK, 2003; BERENTE, 2004; www.statsoft.com/textbook/stathome.html).

Die Durchführung des knn-Vergleichs ist denkbar einfach. Die Proben werden im Raum der Originalvariablen dargestellt, neue, unbekannte Proben werden dann der Klasse zugeteilt, deren Mitglieder den geringsten räumlichen Abstand zur Probe aufweisen. Als Abstandsmaß dient zumeist die Euklidische Distanz, es sind aber auch andere Abstandsmaße, wie zum Beispiel die Minkowski-Distanz, möglich. Ein wichtiger Parameter der Analyse ist die Anzahl der benachbarten Proben, die zur Klassifizierung herangezogen werden sollen. Da diese Anzahl mit *k* bezeichnet wird, wird die Analyse auch k-nächte Nachbarn (knn) Vergleich genannt (KLIMMEK, 2003; BERENTE, 2004; http://www.statsoft.com/textbook/stathome.html. Version: 2004).

Ein weiterer vorteilhafter Aspekt des knn-Vergleichs ist, dass keine Annahmen über Verteilungen der Variablen, wie etwa Normalverteilung, vorausgesetzt werden. Die Methode kann daher in vielen Fällen angewendet werden, in denen die Voraussetzungen für andere Methoden nicht erfüllt werden können (HENRION und HENRION, 1995).

Lineare Diskriminanzanalyse (LDA). Die Lineare Diskriminanzanalyse ist ein Verfahren, das durch Linearkombination der ursprünglichen Variablen neue kanonische Variablen generiert. Dabei werden nicht relevante Variablen im Vergleich schwächer gewichtet und stören damit weniger bei der Klassifizierung. Trotzdem

ist es vorteilhaft, nur relevante Variablen zur LDA heranzuziehen, da das Modell damit einfacher und die Gefahr einer Übermodellierung geringer wird (Berente, 2004). Bei der LDA wird zwar eine Stichprobe aus einer multivariaten Normalverteilung mit homogenen Varianzen der Variablen vorausgesetzt, eine geringe Verletzung dieser Kriterien wirkt sich jedoch erfahrungsgemäß selten auf das Klassifikationsergebnis aus (Klimmek, 2003).

Ähnlich wie bei der Hauptkomponentenanalyse werden in der LDA optimale Linearkombinationen der gemessenen Variablen gesucht. Der Unterschied zur Hauptkomponentenanalyse liegt allerdings im angestrebten Ziel dieser Neukombination. Während diese versucht, die maximale Varianz in den Faktoren zu erhalten, strebt die LDA nach einer möglichst optimalen Trennung der vorgegebenen Objektklassen. Dies ist eine entscheidende Voraussetzung für eine spätere erfolgreiche Klassifikation von unbekannten Proben (Klimmek, 2003). Da die LDA, wie auch der knn-Vergleich, zu den überwachten Verfahren gezählt wird, ist auch hier eine Überprüfung der Klassifizierung auf overfitting notwendig. Auch bei der LDA ist die leave one out-Methode gut dazu geeignet und wird häufig verwendet (Klimmek, 2003).

Neuronale Netzwerke. Neuronale Netzwerke erfuhren in den letzten Jahren eine stark steigende Bedeutung als Methode zur Lösung von unterschiedlichsten Problemstellungen. Dafür sind mehrere Gründe verantwortlich. So können mit der Hilfe von Neuronalen Netzwerken äußerst komplexe Funktionen nichtlinear modelliert werden. Im Vergleich zu klassischen linearen Modellierungsfunktionen ermöglicht dies eine exaktere Beschreibung von vielen natürlichen Vorgängen. Diese lassen sich sonst nur sehr aufwändig und erst durch die Einführung einer großen Anzahl an Variablen durch lineare Modelle beschreiben. Ein anderer Grund für die fortschreitende Verbreitung von Neuronalen Netzwerken ist ihre einfache Anwendung. Ihre hervorstechende Eigenschaft ist, dass sie in der Lage sind, die passende Modellierungsfunktion selbstständig anhand von Trainingsdaten zu erlernen. Der Anwender braucht dabei nicht über detailliertes Wissen über den Aufbau des Netzes, die Datenaufbereitung oder spezielle Auswertungsmethoden zu verfügen. Das nötige Vorwissen für die erfolgreiche Anwendung eines Neuronalen Netzwerks ist weitaus geringer als das, welches zur Anwendung traditionellerer nichtlinearer statistischer Methoden erforderlich ist
(http://www.statsoft.com/textbook/stathome.html).

Der Aufbau eines einfachen Netzes mit der am häufigsten verwendeten dreistufigen feed-forward-Topologie verfügt über eine Eingabeschicht, eine versteckte Schicht und eine Ausgabeschicht (Abb. 1). Die Funktionsweise des Netzes ist intuitiv erfassbar. Die Neuronen der Eingangsschicht empfangen die Eingaben von einer oder mehreren Quellen (z.B. Messwerte von Sensoren). Jedes einzelne davon vergleicht nun die aufsummierten Eingaben mit einem Schwellenwert, bei dessen Überschreitung das Neuron aktiv wird und nach Anwendung seiner Propagierungsfunktion (auch Aktivierungs- bzw. Transferfunktion) ein Signal abgibt. Die von den Neuronen der Eingangsschicht propagierten Signale wirken als Eingaben auf die Neuronen der versteckten Schicht, die das Signal analog zur Eingangsschicht verarbeitet. Dasselbe gilt auch für die Ausgabeschicht, die die Signale schließlich über die Propagierungsfunktion ihrer Neuronen als Antwort auf die Ursprungsdaten ausgibt.

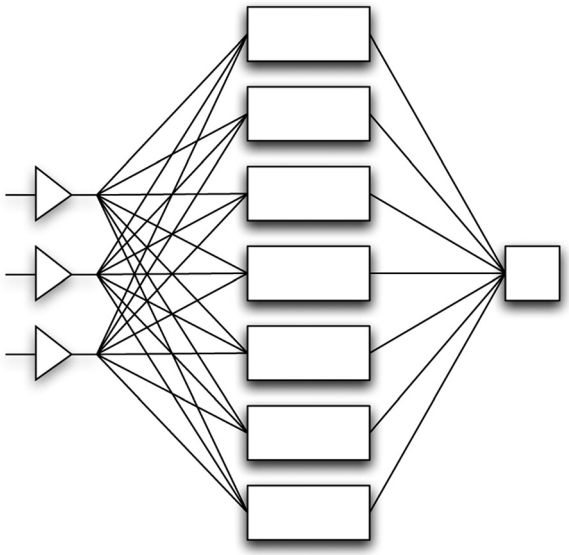

Abb. 1: Topologie eines dreischichtigen, vollständig verbundenen, *feed-forward* Neuronalen Netzes mit drei Eingabeeinheiten, sieben Neuronen in der versteckten Schicht und einem Neuron in der Ausgabeschicht

Die Verwendung von Neuronalen Netzwerken zur Lösung von Klassifizierungsproblemen funktioniert folgendermaßen: Im ersten Schritt wird das Netzwerk mit einem bekannten Datensatz trainiert. Das Netzwerk lernt dabei die Eingabewerte den richtigen Ausgabewerten (z.B. Probenklassen bzw. -gruppen) zuzuordnen. Als Lernalgorithmus dient hierbei in den meisten

Fällen die Rückwärtspropagierung (back propagation). Ist das Neuronale Netzwerk auf diese Art trainiert, kann es zur Klassifizierung unbekannter Proben verwendet werden. Wichtig ist dabei, dass Neuronale Netzwerke nur mit numerischen Werten arbeiten können. Nominale Variablen, wie sie sehr oft bei Klassifizierungsaufgaben vorkommen, müssen in geeigneter Form numerisch kodiert werden. Dies kann etwa durch einfaches Durchnummerieren geschehen, allerdings führt man dadurch zwangsweise eine eventuell nicht den natürlichen Gegebenheiten entsprechende Rangordnung ein. Eine gute Alternative ist es, die Klassen zu der eine Probe nicht gehört, mit 0, alle anderen mit 1 zu bezeichnen
(www.statsoft.com/textbook/stathome.html).
Beim Einsatz der oben beschriebenen Verfahren gibt es auch einige Nachteile zu beachten. So können Neuronale Netzwerke nur in dem Wertebereich eingesetzt werden, in dem sie trainiert wurden. Treten abweichende Eingaben auf, kann das Netzwerk nicht zuverlässig arbeiten, da die in einem bestimmten Wertebereich trainierten Propagierungsfunktionen nicht zu einer Extrapolation fähig sind. Ein weiteres Problem ist die Gefahr eines übertrainierten Neuronalen Netzwerkes, das zwar alle Proben eines Trainingsdatensatzes richtig bewerten kann, bei unbekannten Proben aber versagt. Dieses overfitting, das ein generelles Problem bei überwachten Vorhersageverfahren ist, wurde bereits in den Abschnitten zu knn-Vergleich und Linearer Diskriminanzanalyse behandelt. Auch bei Neuronalen Netzwerken kann dem Problem durch die Verwendung der leave one out-Methode begegnet werden. Zu beachten ist weiters, dass die Rückpropagierung nicht in jedem Fall zu einem optimalen Ergebnis führt, da sie möglicherweise nicht auf das absolute Fehlerminimum, sondern ein lokales Minimum im mehrdimensionalen Fehlerraum zustrebt. Dem wird vor allem durch die Topologie des Netzwerkes und durch die Schrittweite der wiederholten Veränderungen bei der Rückpropagierung begegnet.

Weinproben

Es wurden 59 Weinproben der Sorte 'Grüner Veltliner', Jahrgang 2004, die zum Zeitpunkt der Durchführung dieser Arbeit im Lebensmitteleinzelhandel bzw. bei den Produzenten erhältlich waren, zur Herkunftsbestimmung herangezogen. Inklusive zusätzlich durchgeführter Messwiederholungen wurden insgesamt 70 Analysen vorgenommen. Die Weine kamen aus folgenden österreichischen Weinbaugebieten (in Klammer die Anzahl der Proben): Weinviertel (17), Donauland (15), Neusiedlersee (12), Wachau (6), Kremstal (3), Neusiedlersee-Hügelland (2), Kamptal (2), Traisental (1) und Wien (1). Daraus ist ersichtlich, dass zwar nicht aus allen österreichischen Weinbaugebieten Proben zur Verfügung standen, die Gebiete, in denen der 'Grüne Veltliner' die größte Bedeutung hat, jedoch vertreten waren.

GC-Analytik

Die verwendete Analysenmethode der GC-Messung von flüchtigen sowie nach enzymatischem Abbau glykosidischer Bindungen flüchtigen Weininhaltsstoffen wurde im Zuge einer zweiwöchigen Mitarbeit in der Arbeitsgruppe von Dr. Giuseppe Versini am Istituto Agrario San Michele all'Adige, Italien, studiert und übernommen. Die Methode war zum Zeitpunkt des Aufenthalts im routinemäßigen Einsatz zur Bestimmung von bis zu etwa 70 verschiedenen Aromakomponenten und deren Präkursoren, wie glykosidisch gebundenen Terpenen und anderen Verbindungen.
Das Prinzip ist eine Festphasenextraktion (Solid Phase Extraction, SPE) der flüchtigen und der glykosidisch gebundenen Aromastoffe mit Hilfe eines unpolaren Polystyrol-Divinylbenzol Copolymers (Isolute ENV+, Fa. Separtis, D-79639 Grenzach-Wyhlen) und anschließender Trennung der beiden Fraktionen durch selektive Elution mit Lösungsmitteln unterschiedlicher Polarität (Dichlormethan bzw. Methanol). Die mit Methanol eluierte Phase, die die gebundenen Aromakomponenten enthält, wird anschließend einem enzymatischen Abbau der glykosidischen Bindungen und einer nachfolgenden Flüssig-flüssig-Extraktion mit unpolarem Lösungsmittel unterzogen. Nach einer Konzentration werden die Extrakte mittels Gaschromatographie an einer polaren Säule aufgetrennt und die Einzelkomponenten mit einem Flammenionisationsdetektor (FID) gemessen.
Die Identifikation der einzelnen Komponenten ist bei dieser Methode, im Gegensatz zu anderen, wie etwa einer gaschromatographischen Trennung mit massenspektroskopischer Detektion (GC-MS), ausschließlich über den Vergleich der Retentionszeiten mit Aromastandards möglich. Die Empfindlichkeit der Methode ist allerdings sehr hoch. Die Nachweisgrenzen für einzelne Komponenten liegen im Idealfall in einem Konzentrationsbereich von weniger als 5 µg/l (Versini, pers. Mitt., 2004). Die Konzentrationsbestimmung erfolgt auf halbquantitative Art über den Vergleich mit einem internen Standard bekannter Konzentration. Durch den relativ unspezifisch arbeitenden FID kann auf eine Berücksichtigung unterschiedlicher Signalstär-

ken der verschiedenen detektierten Komponenten verzichtet werden, da diese Unterschiede im Normalfall kleiner als die maximale Genauigkeit der Methode ausfallen. Ausgenommen davon sind die im Extrakt glykosidisch gebundener Aromakomponenten enthaltenen Diole, was aber nicht in Zusammenhang mit der Detektionsmethode steht. In San Michele all'Adige wurde bezüglich der Diole die Erfahrung gemacht, dass bei der Flüssig-flüssig-Extraktion nach dem enzymatischen Abbau nur etwa 50% ihrer Gesamtmenge erfasst werden. Daher werden die gemessenen Peakflächen bei den Diolen vor den folgenden Auswertungsschritten verdoppelt (VERSINI, pers. Mitt. 2004).

Die am Istituto Agrario San Michele all'Adige verwendete Analysenmethode wurde an der Höheren Bundeslehranstalt und Bundesamt für Wein- und Obstbau weitestgehend identisch übernommen. Auf Grund von unterschiedlichen Geräteausstattungen und abweichenden Gegebenheiten im Labor waren geringfügige Adaptionen notwendig.

GC-Parameter

Zur Analyse wurde ein Gaschromatograph HP 5890 Series II der Firma Agilent Technologies (Palo Alto, Kalifornien/USA) verwendet. Als Säule diente eine polare Polyethylenglykol (PEG)-Säule (INNOWAX, 30m * 0,32 mm * 0,5 µm, Fa. J&W Scientific). Zur Detektion diente ein Flammenionisationsdetektor (FID).

Die Injektion der Proben in die GC-Apparatur erfolgte anfänglich manuell mit Hilfe einer 10 µl Autosampler-Injektionsspritze (Fa. SGE 10F-4.2/0.63C). Bei Proben freier Aromasubstanzen wurde jeweils genau 1 µl injiziert, bei Proben gebundener Substanzen wurde die Injektionsmenge auf Grund der geringeren Aromakonzentration auf 1,5 µl erhöht. Nach der Analyse von etwa der Hälfte der Probenweine wurde ein Autosampler (HP 7673, Fa. Agilent) installiert, wodurch die Probenaufgabe automatisiert werden konnte. Vom Abschluss der Probenvorbereitung bis zur Injektion in die GC-Apparatur wurden die vorbereiteten Probenextrakte bei einer Temperatur von -18 °C im Gefrierschrank aufbewahrt.

Die Laufbedingungen der GC wurden ebenfalls von der am Istituto Agrario San Michele angewendeten Methode übernommen. Die Retentionszeiten der einzelnen Substanzen waren aber auf Grund der Unterschiede der eingesetzten GC-Apparaturen nicht identisch, sehr wohl aber die Reihenfolge, in der die aufgetrennten Substanzen detektiert werden konnten. Daher wurde eine Bestimmung der Retentionszeiten von 32 bekannten über die gesamte Laufzeit verteilten Aromastandards vorgenommen. In der Folge wurden diese Messwerte dazu verwendet, um mit Hilfe von identischen, sowohl in San Michele als auch in Klosterneuburg gemessenen Weinproben, eine Liste von Retentionszeiten von 176 bekannten und unbekannten flüchtigen Komponenten zu erstellen. Da die eindeutige Identifikation aller Aromen nicht Ziel dieser Arbeit war, wurden die unbekannten Substanzen einfach zusammen mit den bekannten mit einer fortlaufenden Nummer versehen und nachfolgend über diese referenziert.

Im Verlauf der vorbereitenden Arbeiten wurde der Flussgasdruck am Säulenanfang soweit variiert, dass sich ein optimales Verhältnis von Auflösung und Peakbreite ergab. Die restlichen Laufbedingungen blieben dabei unverändert und wurden wie in Tabelle 1 angegeben eingestellt. Da die vorhandene GC-Apparatur keine Möglichkeit bot, die Druckverhältnisse im Verlauf eines GC-Laufes temperaturabhängig zu regeln, wurden die Werte für eine Temperatur von 100 °C berechnet (HP Flow Calc Version A.02.07). Das bedeutet, dass die tatsächlichen Gasflussbedingungen innerhalb des Gaschromatographen über weite Teile eines Analysenlaufes unbekannt waren und es daher notwendig war, die einmal gewählten Einstellungen bei allen Analysen beizubehalten, um die Vergleichbarkeit zu gewährleisten.

Das in Tabelle 2 dargestellte Temperaturprofil eines Laufes setzte sich aus vier Temperaturstufen zusammen, wovon die letzte nicht mehr der Analyse, sondern der Säulenreinigung nach einem Lauf diente.

Tab. 1: GC-Laufbedingungen

H_2-Gasfluss (100°C)	30 ml/min
H_2-Säulenfluss (100°C)	1,26 ml/min
Splitverhältnis (100°C)	1:24
Druck am Säulenanfang	30 kPa
Injektortemperatur	250°C
Detektortemperatur	250°C

Datenerfassung

Tab. 2: Temperaturprogramm eines GC-Laufs

Temperaturerhöhung °C/min	Haltetemperatur °C	Haltezeit min
-	40	4
2,5	175	10
10	190	40
10	191	30

Die Erfassung und Aufzeichnung der während der GC-Läufe gewonnenen Messdaten erfolgte computerunterstützt mit Hilfe der Software GC ChemStation Rev. A.09.03(1417) von Agilent Technologies, die auch die Steuerung der gesamten GC-Apparatur übernahm.
Nach der Erfassung der Peakflächen der Aromakomponenten wurden diese mit Hilfe der Peakfläche des internes Standards (440 µg/l 1-Heptanol) in Ausgangskonzentrationen in der jeweiligen Weinprobe umgerechnet. Auf eine Berücksichtigung der leicht differierenden FID-Responses der verschiedenen Komponenten wurde dabei, wie bereits zuvor erwähnt, verzichtet. Nur bei den Diolen in der Fraktion der gebundenen Komponenten wurden die ebenfalls oben erwähnten Extraktionsverluste durch eine Multiplikation mit zwei ausgeglichen. Die Errechnung der Konzentration in den Proben erfolgte nach folgender Formel:

$$C_i = \frac{Ai \,.\, C_{IS}}{A_{IS}}$$

C_i Konzentration der jeweiligen Aromakomponente
A_i Fläche des Komponentenpeaks
C_{IS} Konzentration des internen Standards
A_{IS} Peakfläche des internen Standards

Aufbereitung der Messdaten

Die Vorbereitung und Konvertierung der GC-Messdaten erfolgte mit Hilfe des freien Softwarepakets OpenOffice.org. Zur statistischen Auswertung der Daten diente die ebenfalls freie Statistiksoftware R.
Vor der eigentlichen Auswertung mit dem Ziel der Herkunftserkennung wurden die Messdaten aufbereitet sowie die für die Fragestellung der Herkunftserkennung signifikanten Daten selektioniert und zusammengefasst. Fehlende Messwerte (in der jeweiligen Probe nicht nachweisbare Aromakomponenten) wurden durch den Wert 0 ersetzt, Flächen von überlappenden Komponentenpeaks wurden zusammengefasst und in der Folge gemeinsam behandelt. Nachdem die Rohdaten damit für die anschließenden Data Mining-Prozesse vorbereitet waren, wurde versucht die relevanten Daten herauszufiltern. Das heißt, es wurden die Aromakomponenten gesucht, die einen zur Gebietsidentifikation beitragenden Informationsgehalt aufwiesen. Die dazu verwendete Methode war der varianzanalytische F-Test. Durch eine Hauptkomponentenanalyse wurde zusätzlich versucht, die in den Messwerten enthaltenen Informationen über das Herkunftsgebiet in einigen wenigen Faktoren zusammenzufassen, um sie besser handhabbar zu machen. Die Hauptkomponentenanalyse erfolgte über die Berechnung der Eigenvektoren der Kovarianzmatrix aus den Datensätzen und zusätzlich über die Eigenvektoren ihrer Korrelationsmatrix. Die Kovarianzmatrix ist ein Maß für die Varianzen aller Messwerte einer Probe und damit die multidimensionale Entsprechung der einfachen Varianz, die im Gegensatz dazu nur ein einzelnes Merkmal, also nur eine Dimension, beschreibt. Die Korrelationsmatrix drückt die Korrelation zwischen den ursprünglich gemessenen einzelnen Faktoren, also den Konzentrationen der signifikanten Aromakomponenten, in Zahlen zwischen -1 (negativ lineare Korrelation), 0 (keinerlei Korrelation) und +1 (positiv lineare Korrelation), aus. Sie wurde von der verwendeten Statistiksoftware R mit Hilfe der Pearson-Korrelation gebildet.
Bei dieser Methode werden die Kovarianzen der Messwerte durch die Produkte ihrer Standardabweichungen dividiert und damit sowohl standardisiert als auch skaliert. Je nach Zusammensetzung der Messwerte kann entweder die Verwendung der Eigenvektoren der Korrelationsmatrix oder der Kovarianzmatrix zu besseren Ergebnissen bei der Hauptkomponentenanalyse führen (www.statsoft.com/textbook/stathome.html). Da nicht bekannt war, wie dies im Fall der vorliegenden Daten aussah, wurden beide Varianten berechnet und alle Ergebnisse für die anschließenden Vorhersagen verwendet.

Vorhersage der Herkunftsgebiete

Zur Herkunftsvorhersage dienten die bereits beschriebenen Verfahren des knn-Vergleichs, der Linearen Diskriminanzanalyse und der Einsatz eines Neuronalen Netzwerks. Zur Auswertung wurden aus den Ursprungsdaten mehrere Teildatensätze gebildet und mit Hilfe der einzelnen Methoden bewertet. So wurde auf Grund der Tatsache, dass von nur drei Weinbaugebieten (Donauland, Neusiedlersee, Weinviertel) eine größere Probenanzahl zur Verfügung stand, eine separate Vorhersage mit den Analysenergebnissen ausschließlich dieser Proben durchgeführt.
Bei der Methode der k-nächsten Nachbarn wurde die Software R dazu eingesetzt, die vorhandenen Datensätze zu beurteilen, um die idealen Parameter für den folgenden knn-Vergleich herauszufinden. Dazu wurde die integrierte Trainingsfunktion verwendet, die einen gegebenen Datensatz von Proben bekannter Gruppenzugehörigkeit analysiert und daraus die Parameter ableiten kann, die bei einem knn-Vergleich zur optimalen Klassifizierung der Proben eben dieses Datensatzes

Tab. 3: Binärcodierung der Herkunftsgebiete für die Vorhersage mit Hilfe des Neuronalen Netzwerks

Anbaugebiet	Donau-land	Kamptal	Kremstal	Neusiedler-see	Neusiedler see-Hügelland	Traisen-tal	Wachau	Wein-viertel	Wien
Donauland	1	0	0	0	0	0	0	0	0
Kamptal	0	1	0	0	0	0	0	0	0
Kremstal	0	0	1	0	0	0	0	0	0
Neusiedlersee	0	0	0	1	0	0	0	0	0
Neusiedlersee-Hügelland	0	0	0	0	1	0	0	0	0
Traisental	0	0	0	0	0	1	0	0	0
Wachau	0	0	0	0	0	0	1	0	0
Weinviertel	0	0	0	0	0	0	0	1	0
Wien	0	0	0	0	0	0	0	0	1

führen würden. Die so ermittelten Parameter wurden nachfolgend zur tatsächlichen Herkunftsvorhersage mit Hilfe der leave one out-Methode verwendet.

Bei der LDA werden, ähnlich wie bei der Hauptkomponentenanalyse, neue kanonische Variablen aus den Datensätzen gebildet. Ein Vorteil dabei ist, dass Variablen, die weniger Informationsgehalt aufweisen, schwächer gewichtet werden. Aus diesem Grund eignet sie sich auch gut zur Auswertung von Rohdaten ohne jegliche Vorselektion, obwohl eine solche auch bei der LDA zu besseren Ergebnissen führen kann. Da die LDA selbstständig eine Zusammenfassung der Probeninformation vornimmt, wurden hier die mittels Hauptkomponentenanalyse gewonnenen zusammengefassten Daten nicht zur Vorhersage verwendet. Stattdessen wurde sie ausschließlich aus den Daten über die Aromakonzentrationen aller Proben beziehungsweise der Probenteilmenge aus den Weinbaugebieten Donauland, Neusiedlersee und Weinviertel vorgenommen.

Zur Vorhersage der Weinbaugebiete mit Hilfe eines Neuronalen Netzwerks wurde ebenfalls das Statistikpaket R verwendet. Sie erfolgte durch ein vollständig verbundenes feed-forward-Netz mit einer einzelnen, zehn Neuronen enthaltenden, versteckten Schicht. Die Anzahl der Neuronen der Eingabeschicht wurde durch die Anzahl der einzugebenden Messwerte aus den Probendatensätzen vorgegeben. Verwendet wurden hierfür die Ergebnisse der Hauptkomponentenanalyse, also die jeweils ersten 15 extrahierten Vektoren eines jeden der zur Verfügung stehenden Datensätze. Es ergaben sich daher 15 Neuronen auf der Eingabeseite. Die Anzahl der Neuronen auf der Ausgabeseite war durch die Anzahl der Weinbaugebiete, denen die Proben zugeordnet werden sollten, vorgegeben. Sie umfasste neun Neuronen und wurde in dieser Form auch bei der Auswertung der ausschließlich aus den Weinbaugebieten Donauland, Neusiedlersee und Weinviertel stammenden Teilmenge der Datensätze verwendet. Dies diente vor allem dazu, um durch die Verwendung eines identischen Neuronalen Netzwerks die Ergebnisse besser vergleichbar zu machen. Zuletzt mussten noch die Abbruchbedingungen zur Beendigung der Trainingssequenz festgelegt werden. Die Parameter wurden so gewählt, dass das Training beendet wurde, wenn eine der folgenden beiden Bedingungen eintrat:

maximale Iterationszahl : 1000
minimale Fehlerverringerung/Iteration: 0,01

Vor dem Training und der Auswertung musste die Herkunft der Proben noch in Zahlen gefasst werden, da Neuronale Netzwerke nicht mit nominalen Werten umgehen können. Dazu wurde eine binäre Kodierung des Herkunftsgebietes gewählt, wie sie in Tabelle 2 dargestellt ist, und das Netzwerk in der Folge trainiert. Auch hier wurde wieder die leave one out-Methode verwendet. Das Training eines Neuronalen Netzwerkes läuft so ab, dass die Eingabewerte Probe für Probe in die Eingabeschicht gespeist und die Propagierungsfunktionen innerhalb des Netzwerks solange adaptiert werden, bis das Signal in der Ausgangsschicht genau dem Erwartungswert entspricht oder eine der oben genannten Abbruchbedingungen eintritt. Als Erwartungswert diente in diesem Fall das binär kodierte Herkunftsgebiet der jeweiligen Weinproben. Die einzelnen Neuronen der Ausgangsschicht sollten daher im Idealfall einmal den Wert 1 und ansonsten nur den Wert 0 liefern. In der Praxis ist dieses Ergebnis im Normalfall allerdings nur näherungsweise erreichbar. Nachdem das Netzwerk trainiert war, wurden die bis dahin ausgesparten Proben eingespeist und das resultierende Signal der Ausgangsschicht aufgezeichnet. Diese Prozedur wurde mit jeder Probe eines Datensatzes wiederholt.

Ergebnisse und Diskussion

GC-Analysen und Datenerfassung

Bei der Erfassung der GC-Messdaten traten einige Schwierigkeiten bei der Wiedererkennung einzelner Peaks auf, was vor allem darauf zurückzuführen war, dass eine große Anzahl von Peaks sehr nahe beieinander lag. Kleinere Verschiebungen der Laufgeschwindigkeit, wie sie vor allem von Woche zu Woche auftraten, verlangten nach einer wiederholten manuellen Anpassung der in der Software gespeicherten Retentionszeiten aller gemessenen Komponenten. Bei je 176 gemessenen Peaks nahm dies einen nicht zu vernachlässigenden Teil des Arbeitsaufwandes für die Datenerfassung in Anspruch. Die Verschiebungen der Retentionszeiten waren dabei zufällig und nicht vorhersehbar. Verursacht wurden sie vermutlich hauptsächlich durch geringe Variationen des Drucks im Flussgas zwischen zwei Einschaltperioden. Da der Flussgasstrom innerhalb einer Messetappe aber immer eingeschaltet blieb und nur am Wochenende unterbrochen wurde, kam es dabei zu den beobachteten wöchentlichen Schwankungen. Die Schwankungsbreite bewegte sich in einem relativ geringen Bereich von einigen Sekunden. Durch die teilweise sehr nah beisammen liegenden Komponentenpeaks kam es jedoch immer wieder zu falschen Zuordnungen, die einen manuellen Eingriff notwendig machten. Nach einer entsprechenden Korrektur war es möglich, die 176 der gemessenen Substanzen zuverlässig wieder zu erkennen. Abbildung 2 zeigt ein typisches Chromatogramm so wie es im Zuge der Analysen erhalten wurde. Bei der Analyse der glykosidisch gebundenen Aromakomponenten war deutlich ersichtlich, dass diese in Grüner Veltliner-Weinen nur in äußerst geringen Konzentrationen nachzuweisen waren. Bereits zu diesem Zeitpunkt wurde angenommen, dass sie sich aus diesem Grund nicht zur Herkunftserkennung von Weinen eignen würden, was in der Folge auch bestätigt werden konnte Die Trefferquote lag bei Vorhersagen auf Grundlage dieser Werte nie höher als die Zufallswahrscheinlichkeit. Aus diesem Grund werden die Ergebnisse aus diesen Vorhersagen hier in der Folge nicht weiter betrachtet. Nach der Anpassung der Retentionszeiten konnten die Peaks von der Software zugeordnet und ihre Flächen berechnet werden. Aus diesen Ergebnissen wurde über den Vergleich mit der Peakfläche des n-Heptanol-Standards mit bekannter Konzentration (Peak 63) die Konzentration der einzelnen flüchtigen Komponenten errechnet.

Da der absoluten Genauigkeit der verwendeten Analytik, wie bereits weiter oben angesprochen, auf Grund der halbquantitativen Methode klare Grenzen gesetzt waren, sind die erhaltenen Ergebnisse nur innerhalb der durchgeführten Probenserie uneingeschränkt vergleichbar. Um zu überprüfen, inwieweit diese Vergleichbarkeit durch auftretende Schwankungen der Messwerte beeinflusst wurde, wurden einzelne Proben mehrfach analysiert und die Ergebnisse verglichen. Dazu wurden die Mittelwerte und Standardabweichungen der einzelnen gemessenen Aromakomponenten herangezogen und letztere in prozentuellen Anteilen am jeweiligen Mittelwert ausgedrückt. Es ist klar, dass die Abweichungen bei in sehr geringen Konzentrationen (im Bereich von <5 µg/l) auftretenden Komponenten weit höher waren als bei den anderen. Hier wurde

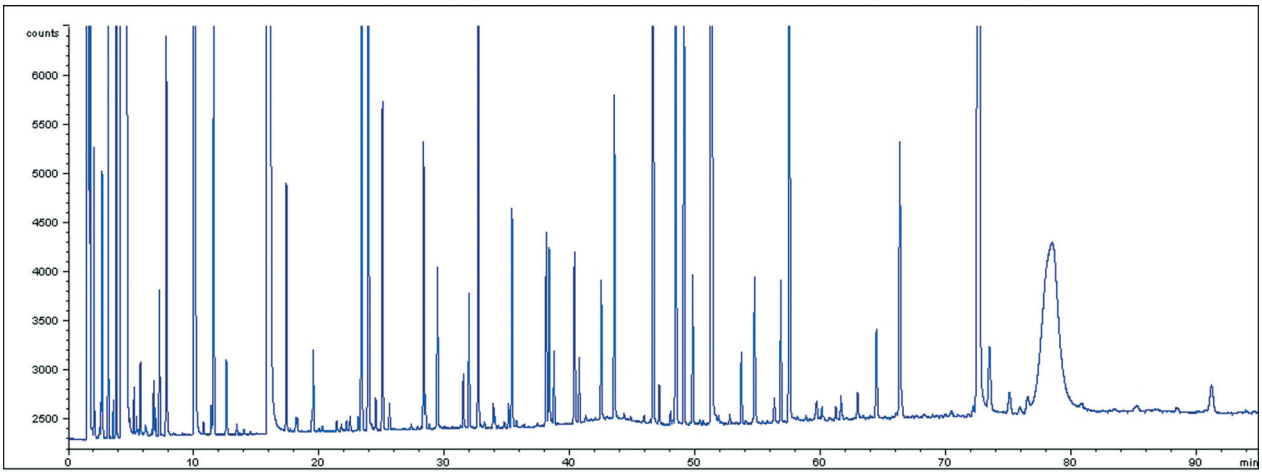

Abb. 2: Typisches Gaschromatogramm eines Weins der Sorte 'Grüner Veltliner'

die Erfahrung gemacht, dass allein durch eine kleine Variabilität bei der Begrenzung der Peaks im Zuge der Integration Unterschiede im Bereich von über 25% auftraten. Bei Komponenten mit Konzentrationen von 5 µg/l lagen die Standardabweichungen größtenteils im Bereich von weniger als 15 oder sogar 10% des Mittelwerts der Messungen. Messungen von Komponenten mit Konzentrationen von ca. 20 bis 50 µg/l aufwärts führten oft zu Standardabweichungen von deutlich weniger als 5% des Messwertes, was bereits einer sehr guten Wiederholbarkeit entspricht.

Statistische Auswertung und Herkunftsbestimmung

Zur statistischen Auswertung der Daten wurden die analytischen Ergebnisse der freien und gebundenen Aromakomponenten getrennt voneinander betrachtet. Es zeigte sich im Verlauf der Auswertungen, dass bei den Gebieten mit einer geringen Probenanzahl, wie bereits anfänglich vermutet, keine zuverlässige Vorhersage der Herkunft möglich war. So stand beispielsweise nur eine einzige Probe aus dem Weinbaugebiet Wien zur Verfügung. Das machte es somit unmöglich, dieses Gebiet mittels eines statistischen Vorhersageverfahrens unter Zuhilfenahme der leave one out-Methode vorherzusagen. Um dieses Problem zu umgehen, wurde eine Teilmenge der Proben gebildet, die ausschließlich aus den Weinbaugebieten Donauland, Neusiedlersee und Weinviertel stammten. Von jedem dieser Gebiete standen mehr als zehn Weine zur Verfügung. Diese zusammen genommen 44 Proben wurden bei allen Auswertungen zusätzlich getrennt von den anderen betrachtet.

Selektion relevanter Daten (F-Test). Der F-Test ergab 41 freie flüchtige Inhaltsstoffe, die eine für die Herkunft signifikante Varianz aufwiesen (Signifikanzniveau: 95%). Tabelle 4 listet alle entsprechenden Komponenten zusammen mit der Nummer, mit der diese während der Analysen bezeichnet wurden, und den entsprechenden Retentionszeiten auf. Dieser Test war allerdings nicht dazu geeignet herauszufinden, zwischen welchen beiden Gebieten die signifikanten Unterschiede bestanden. Auf diese Information wurde daher verzichtet. Für die weiterfolgenden statistischen Tests war sie ohnehin nicht von Bedeutung. Abbildung 3 veranschaulicht am Beispiel von vier Komponenten die Unterschiede von signifikanten und nicht signifikanten Aromakomponenten. Auffällig bei diesen Ergebnissen ist vor allem, dass die Terpene nur eine untergeordnete Rolle zu spielen scheinen. Allerdings kann natürlich nicht ausgeschlossen werden, dass es sich bei einer oder mehreren der nicht identifizierten Komponenten um solche handelte. Allerdings waren die Retentionszeiten der wichtigsten Vertreter der Terpene sehr wohl bekannt und wurden auch nachgewiesen. Dabei wurden jedoch in jedem Fall nur sehr geringe Konzentrationen gefunden, die noch dazu beim F-Test als nicht signifikant herausfielen. Bei den signifkanten Peaks der freien Aromakomponenten waren die bekannten Substanzen deutlich in der Überzahl. Daher ist auch erkennbar, dass es sich dabei hauptsächlich um bei der Gärung gebildete sekundäre Aromen nach der Einteilung von CLARKE und BAKKER (2004) beziehungs-

Tab. 4: Für die Herkunftsidentifikation signifikante freie Aromakomponenten

Nr.	Komponente	Retentionszeit (min.)
1	Isobutylacetat	6,89
3	unbekannt	7,23
7	unbekannt	9,28
9 & 10	unbekannt & Isobutanol	9,98 & 10,01
15	Isoamylacetat	11,56
16	n-Butanol	12,57
19	unbekannt	13,60
25	Capronsäureethylester	17,53
28	n-Pentanol	18,22
33	Hexylacetat	19,49
41	Ethyllactat	23,35
47	3-Ethoxypropanol	25,02
48	cis-3-Hexenol	25,57
58	Caprylsäureethylester	28,30
61	Essigsäure	28,76
79 & 80	unbekannt & ein Butandiol	33,87 & 34
84	unbekannt	35,07
86	Isobuttersäure	35,35
87	unbekannt	35,72
90	unbekannt	37,42
94	Buttersäure	38,32
95	Caprinsäureethylester	38,71
99	Diethylsuccinat	40,71
103	unbekannt	41,99
105	3-Methyl-Thiopropanol	42,47
113	Citronellol	44,83
114	unbekannt	45,37
118	β-Phenylethylacetat	47,10
120 & 121	Geraniol & Capronsäure	48,35 & 48,42
129	β-Phenylalkohol	51,27
141	2-Ethyl-Hexanolsäure	54,71
145	Diethylmalat	56,81
146	Caprylsäure	57,48
151	ein Aminosäureester	59,67
158	4-Vinylguajakol	62,93
160	4-Carboethoxy-γ-Butyrolacton	64,43
167	unbekannt	68,94
168	unbekannt	69,25
170	unbekannt	69,95
173	unbekannt	71,32
174	unbekannt	72,16

weise Fermentationsaromen nach der Einteilung von Rapp (1998) handelt. Dies verwundert auf den ersten Blick ein wenig, da die Fermentationsaromen nicht direkt von der Traube stammen und die Herkunft daher einen weit geringeren Einfluss auf ihre Bildung hat als zum Beispiel die bei der Vergärung eingesetzte Hefe. So basierten beispielsweise die von Rapp et al. (1985) durchgeführten Untersuchungen über die Herkunftsbestimmung von Riesling-Weinen ausschließlich auf der Konzentration von zwölf Monoterpenen. Arrhenius et al. (1996) beachteten jedoch bei ihrer Arbeit über die Charakterisierung von kalifornischen Chardonnay-Weinen auch Fermentationsaromen, obwohl auch sie den Terpenen einen hohen Stellenwert zuwiesen. Nun

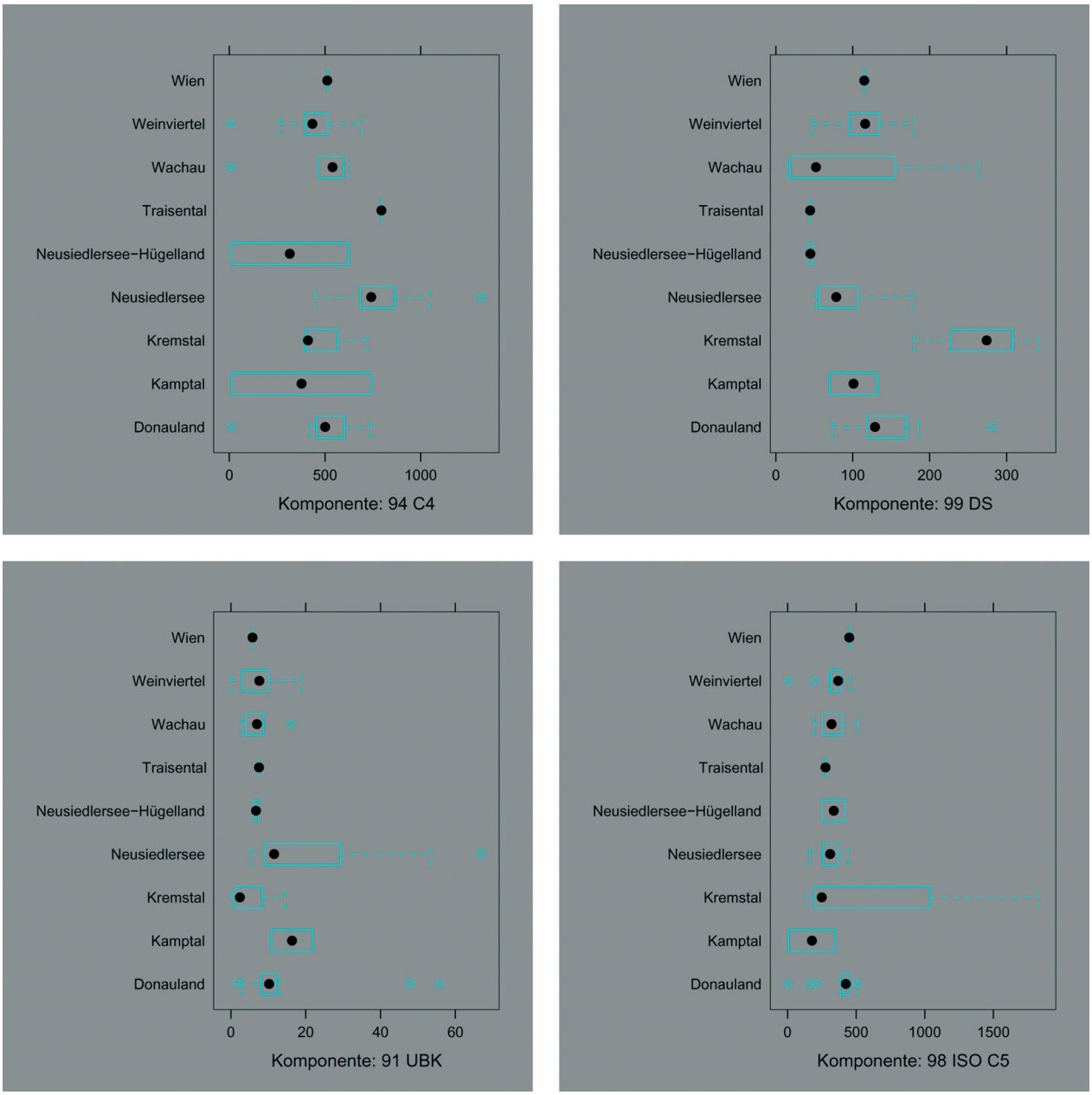

Abb. 3: Verteilung der Konzentrationen von vier Aromakomponenten in verschiedenen Weinbaugebieten. Oben: signifikante Unterschiede zwischen einzelnen Weinbaugebieten; Unten: keine signifikanten Unterschiede (Signifikanzniveau 95%, Konzentrationsangaben auf der x-Achse in µg/l).

wurde die Sorte 'Güner Veltliner' noch nie mit hohen Terpengehalten in Verbindung gebracht. Zusätzlich könnte die relativ kühle Witterung des Jahrgangs 2004 zu einer verringerten Bildung in den Trauben geführt haben, obwohl Grüne Veltliner-Weine dieses Jahrgangs im Allgemeinen durchaus als typisch angesehen wurden und werden. Es musste also versucht werden, die Herkunft mit Hilfe der Analyse von Fermentationsaromen zu bestimmen, wofür es durchaus Vorbilder gibt. In der oben erwähnten Arbeit von ARRHENIUS et al. (1996) konnten auch aus den Fermentationsaromen Herkunftsinformationen abgeleitet werden, wenn auch mit einer im Vergleich zu den Terpenen geringeren Sicherheit.

Zusammenfassung signifikanter Daten (Hauptkomponentenanalyse). Bei den zwei Varianten der Hauptkomponentenanalyse ergab sich ein grundsätzlicher Unterschied in der Beladung der extrahierten Faktoren. Bei der Hauptkomponentenanalyse mit Hilfe der Kovarianzmatrix wurde beinahe die gesamte Varianz auf die ersten beiden Faktoren geladen, während sie bei der Variante auf Basis der Korrelationsmatrix weniger stark zusammengefasst wurde. Dies ist sehr gut in Abbildung 4 sichtbar, die die Beladungen der ersten zehn Faktoren beider Versionen darstellt. Um einen ersten Überblick zu erhalten, wie gut die extrahierten Hauptkomponenten in der Lage waren, die vorhandenen Proben nach dem Weinbaugebiet zu unterscheiden, wurden Diagramme erstellt, die die erste und damit am stärksten beladene Hauptkomponente in einem zweidimensionalen Koordinatensystem der zweiten Hauptkomponente gegenüberstellen. Die Ergebnisse dieser Gegenüberstellung sind in Abbildung 5 zu sehen. In der oberen Reihe sind die Ergebnisse der Hauptkomponentenanalyse der Konzentrationen aller Aromasubstanzen dargestellt. Das linke Diagramm zeigt die Ergebnisse der Hauptkomponentenanalyse mit Hilfe der Eigenvektoren der Kovarianzmatrix, beim rechten Diagramm wurde die Korrelationsmatrix verwendet. Die Diagramme zeigen, dass einige der Weinbaugebiete deutlich voneinander getrennt werden konnten, während dies bei anderen nicht in demselben Maß möglich war. Ersichtlich ist auch, dass die Trennung bei der Variante aus den Eigenvektoren der Korrelationsmatrix besser zu sein scheint als bei der Variante auf Basis der Kovarianzmatrix. In der unteren Reihe befinden sich die Ergebnisse der Hauptkomponentenanalyse der Proben, die ausschließlich aus den Weinbaugebieten Donauland, Neusiedlersee und Weinviertel stammten. Die Ergebnisse ähneln bezüglich der Trennung nach den Weinbaugebieten größtenteils denen aus der Analyse der Aromasubstanzen aller Proben. So ist auch hier die Kovarianzmatrix-Variante schlechter aufgetrennt als die Korrelationsmatrix-Variante. Letztere liefert hierbei das eindeutig beste Ergebnis aller durchgeführten Vergleiche und vermag praktisch eindeutig zwischen den

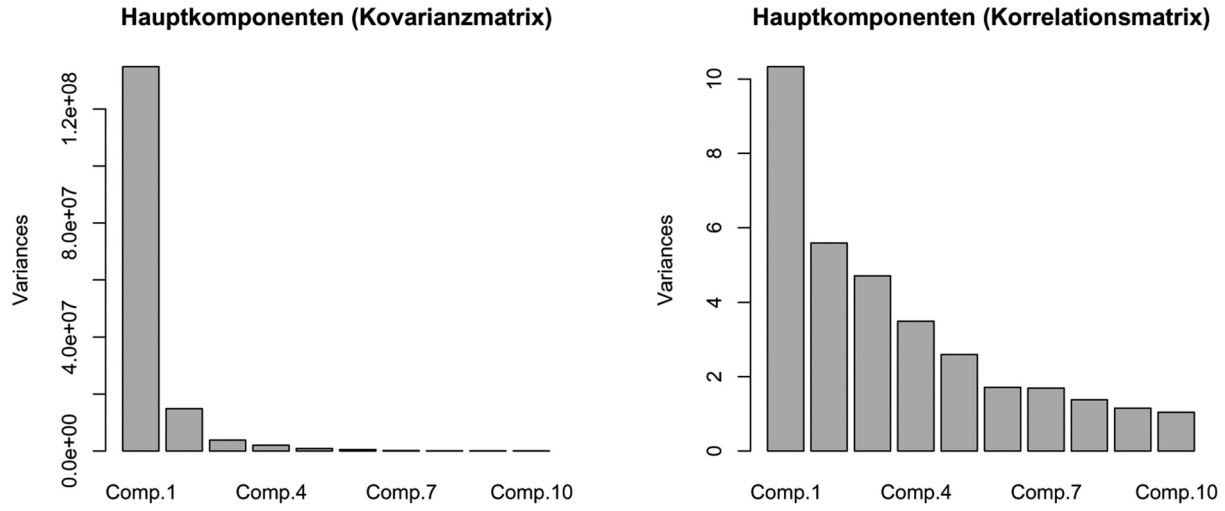

Abb. 4: Faktorladungen der beiden Varianten der Hauptkomponentenanalyse

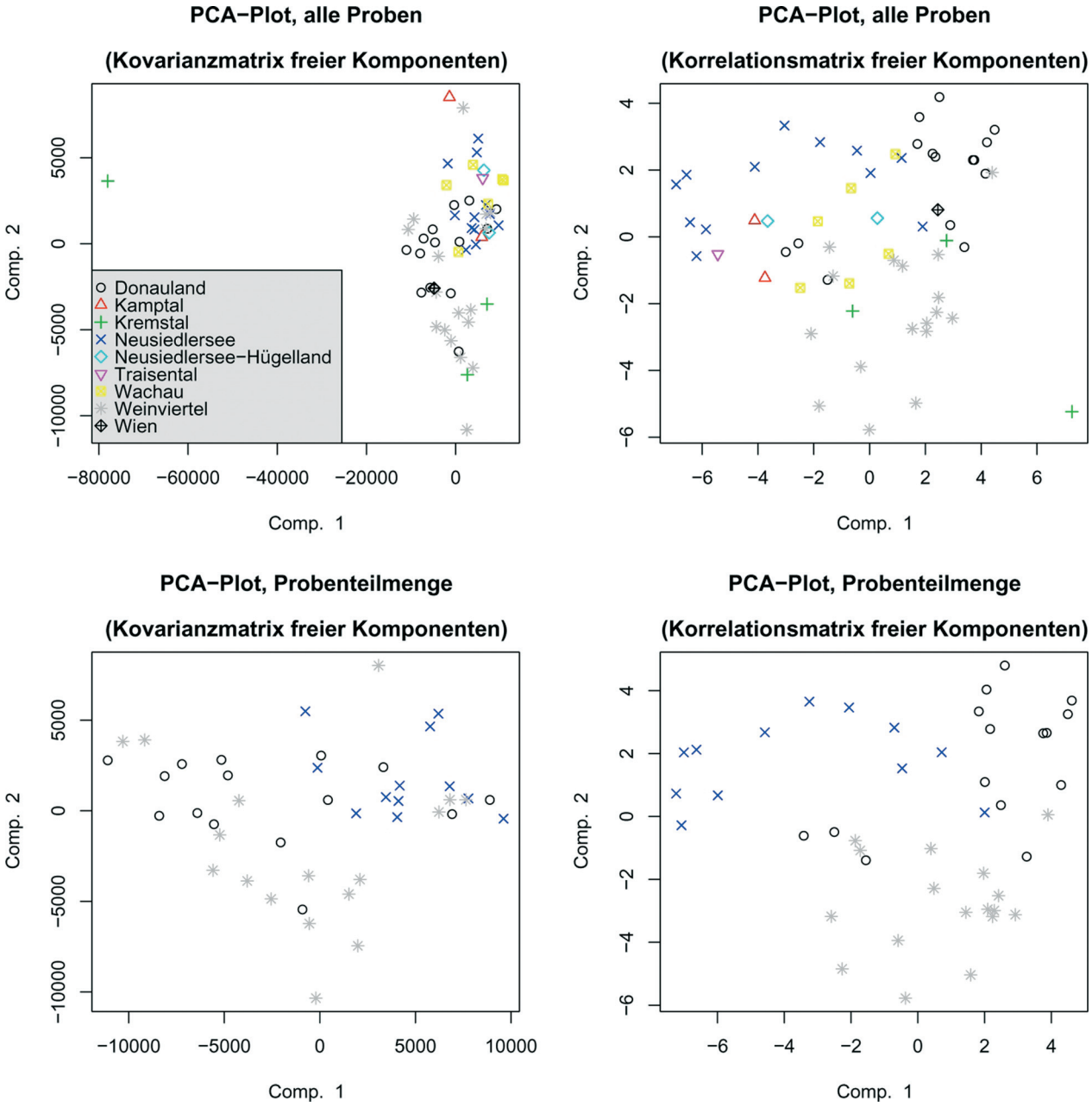

Abb. 5: Gegenüberstellung der ersten beiden Hauptkomponenten aus der Hauptkomponentenanalyse aller Proben sowie der Teilmenge aus den Weinbaugebieten Donauland, Neusiedlersee und Weinviertel

Weinbaugebieten Neusiedlersee und Weinviertel zu unterscheiden. Das Weinbaugebiet Donauland zeigt Überschneidungen, vor allem mit dem Gebiet Weinviertel. Dies ist aber auch nicht weiter verwunderlich, da einige der untersuchten Proben aus Langenzersdorf stammten, das direkt an das Weinbaugebiet Weinviertel grenzt. Aber auch in der Kovarianzmatrix-Variante lassen sich eindeutige Gruppenbildungen zwischen den einzelnen Proben feststellen.

Nach der Hauptkomponentenanalyse wurden die extrahierten Faktoren zur Vorhersage der Weinbaugebiete mit Hilfe des knn-Vergleichs und unter Verwendung eines Neuronalen Netzwerkes verwendet. Dazu wurden die ersten 15 der extrahierten Hauptkomponenten sepa-

rat gespeichert und für die weiteren Analysen verwendet. Diese Zahl wurde gewählt, weil die ersten 15 der extrahierten Faktoren aus der Analyse mit Hilfe der Korrelationsmatrix zusammengenommen gerade mehr als 95% der gesamten Probenvarianz aufnahmen. Bei der Auswertung mit Hilfe der Kovarianzmatrix konnte diese Beladung zwar bereits mit Hilfe der ersten beiden Hauptkomponenten realisiert werden, zur einheitlichen Durchführung der folgenden Auswertungen wurden aber auch hier die ersten 15 Vektoren weiter verwendet. Insgesamt ergaben sich durch die Selektion und Zusammenfassung der Messwerte acht unterschiedliche Datensätze, die für die folgenden Herkunftsvorhersagen verwendet wurden. Zur besseren Übersicht sind diese in Tabelle 5 angeführt und werden in der Folge mit der hier angegebenen Datensatznummer referenziert.

Herkunftsvorhersage mit knn-Vergleich. Die Beurteilung der Datensätze zur Findung der optimalen beim knn-Vergleich anzuwendenden Parameter ergab, dass nur bei zwei Varianten minimale Klassifizierungsfehler von weniger als 40% vorhergesagt wurden. Dies waren die Konzentrationen der signifikanten Aromakomponenten (Datensatz 2, 39% minimaler Klassifizierungsfehler) und die Kombinationsvektoren aus der Hauptkomponentenanalyse mit Hilfe der Korrelationsmatrix (Datensatz 3, 36%). Bei der Teilmenge aus den Proben aus den Weinbaugebieten Donauland, Neusiedlersee und Weinviertel fielen diese Werte mit 16 (Datensatz 2a) bzw. 18% (Datensatz 3a) deutlich besser aus. Für die Vorhersage der Herkunft mittels der leave one out-Methode wurde in der Folge die von der Trainingsfunktion der Statistiksoftware empfohlene Anzahl nächster Nachbarn für die Klassifizierung des jeweiligen Datensatzes verwendet. Tabelle 6 listet die Datensätze und die Anzahl der berücksichtigten Nachbarn auf, die über die Messung der Minkowski-Distanz ermittelt wurde. Die tatsächlich bei der Vorhersage erzielten Trefferquoten sind ebenfalls in dieser Tabelle angegeben.

Der knn-Vergleich lieferte Zugehörigkeitswahrscheinlichkeiten zu den verschiedenen Weinbaugebieten, die sich insgesamt zu 1 ergänzen. Die Proben wurden dem Gebiet, das die höchste Wahrscheinlichkeit aufwies, zugeordnet. Teilweise kam es jedoch zu einer gleich wahrscheinlichen Zuordnung zu zwei verschiedenen Herkunftsgebieten. In solchen Fällen wurde die Vorhersage als nicht richtig gewertet. Auffällig ist vor allem die weit bessere Trefferquote bei den Proben aus Weinbaugebieten, von denen eine größere Probenanzahl zur Verfügung stand. So konnten bei der knn-Vorhersage der Datensätze aller Proben nur Weine aus den Gebieten Donauland, Neusiedlersee und Weinviertel, nicht aber aus den anderen Weinbaugebieten, richtig zugeordnet werden. Damit kamen die Trefferquoten von über 50% nur dadurch zustande, dass der knn-Vergleich alle Proben ausschließlich den drei am häufigsten vorkommenden Weinbaugebieten zuordnete, was die Wahrscheinlichkeit erhöhte, die richtige Herkunft zu erraten. Das und die weit besseren Trefferquoten bei den Datensätzen aus den Gebieten mit einer höheren Probenanzahl lassen erkennen, dass eine Herkunftsvorhersage mit Hilfe des knn-Vergleichs durchaus möglich ist. Es muss allerdings sehr genau darauf geachtet werden, eine größere Anzahl von Proben zu untersuchen, die sich außerdem möglichst gleichmäßig auf alle der in Frage kommenden Herkunftsregionen aufteilen.

Herkunftsvorhersage Lineare Diskriminanzanalyse. Bei der LDA konnten nur Datensätze ausgewertet werden, die nicht zu viele fehlende Aromapeaks (Messert = 0) aufwiesen. Datensatz 2a (alle Aromakomponenten der Proben aus den Gebieten Donauland, Neusiedlersee und Weinviertel) konnte daher nicht zur Herkunftsvorhersage verwendet werden. Das Ergebnis der Klassifizierung mittels Linearer Diskriminanzanalyse war in jedem Fall die Angabe einer einzelnen von den

Tab. 6: Anzahl der zum knn-Vergleich herangezogenen Nachbarn (*k*) und die Ergebnisse der Herkunftsvorhersage

Datensatz Nr.	Beschreibung	Anzahl beachteter Nachbarn	Korrekte Herkunftsvorhersage
2	Konzentrationen aller für die Herkunft signifikanten Aromakomponenten	4	52,63 % (34 von 59)
2a	Teilmenge Weinbaugebiete Donauland, Neusiedlersee und Weinviertel	8	61,36 % (27 von 44)
3	Faktoren der Hauptkomponentenanalyse mit Hilfe der Korrelationsmatrix	5	54,24 % (32 von 59)
3a	Teilmenge Weinbaugebiete Donauland, Neusiedlersee und Weinviertel	7	81,82 % (36 von 44)

zur Auswahl stehenden Klassen. Da keine Zugehörigkeitswahrscheinlichkeiten ausgegeben wurden, kam es bei der LDA auch nie zu dem Fall, dass eine Probe, mit gleicher Wahrscheinlichkeit mehreren Herkunftsgebieten zugeordnet wurde. Die Trefferquote der Klassifizierung lag bei den Datensätzen aus allen Proben (Datensätze 1 und 2), bei nur etwas über 40%. Beim ebenfalls analysierten Datensatz der signifikanten Aromasubstanzen der Proben aus den Weinbaugebieten Donauland, Neusiedlersee und Weinviertel (Datensatz 2a), konnte hingegen eine Trefferquote von knapp 66% erreicht werden. Auffallend ist, dass hierbei im Vergleich zu den knn-Vorhersagen auch Proben richtig klassifiziert wurden, die aus Gebieten stammten, aus denen weniger Weine zur Verfügung standen. So gelang hier sogar in einem Fall die richtige Zuordnung einer von insgesamt nur drei Proben aus dem Weinbaugebiet Kremstal. Tabelle 7 zeigt die bei der Herkunftsvorhersage mit Hilfe der Linearen Diskriminanzanalysen erzielten Ergebnisse.

Tab. 7: Ergebnisse der Herkunftsvorhersage mittels Linearer Diskriminanzanalyse

Datensatz Nr.	Beschreibung	Korrekte Herkunftsvorhersage
1	Konzentrationen aller freien Aromakomponenten	40,68 % (24 von 59)
2	Konzentrationen aller für die Herkunft signifikanten Aromakomponenten	42,37 % (25 von 59)
2a	Teilmenge Weinbaugebiete Donauland, Neusiedlersee und Weinviertel	65,91 % (29 von 44)

Herkunftsvorhersage Neuronales Netzwerk. Es wurden alle bei der Hauptkomponentenanalyse gewonnenen Datensätze zur Vorhersage mit Hilfe des trainierten Neuronalen Netzwerks verwendet. Die Ergebnisse der Auswertung waren eine Matrix von Zahlen zwischen 0 und 1, die die Zuordnung der jeweiligen Proben zu den verschiedenen Weinbaugebieten angaben. Diese von den Ausgabeneuronen des Netzwerkes erhaltenen Werte ergänzten sich nicht notwendigerweise auf eine Summe von exakt 1 und wurden daher so skaliert, dass dies zutraf. Dadurch wurde es möglich, die Wahrscheinlichkeit der Zugehörigkeit zu einem bestimmten Weinbaugebiet in Prozent auszudrücken. Die Proben wurden danach einfach dem Weinbaugebiet zugeordnet, bei dem das Neuronale Netzwerk die größte Zugehörigkeitswahrscheinlichkeit vorhergesagt hatte. Bei allen ausgewerteten Datensätzen lag die Trefferquote bei denen, deren Grundlage die Hauptkomponentenanalyse mit Hilfe der Korrelationsmatrix war, weit höher als bei den auf Grundlage der Kovarianzmatrix ausgewerteten Datensätzen. Dieses Verhalten deckt sich mit den bereits bei den knn-Vergleichen gemachten Beobachtungen. Im Gegensatz zu diesen konnte das Neuronale Netzwerk allerdings auch einige der Proben aus den Gebieten, aus denen nur eine kleinere Probenanzahl untersucht wurde, richtig zuordnen. Eine wirklich hohe Trefferquote von über 77% konnte nur bei der Analyse des Datensatzes 3a (Hauptkomponentenanalyse aus der Korrelationsmatrix der Probenteilmenge) erzielt werden. Dies ist dieselbe Variante, die auch beim knn-Vergleich die höchste richtige Zuordnung von in dem Fall sogar beinahe 82% aufwies. Tabelle 8 zeigt die gesammelten Ergebnisse der Herkunftsvorhersage mit Hilfe des Neuronalen Netzwerks.

Vergleich der Vorhersagemethoden. Von den drei zur Vorhersage der Herkunftsgebiete verwendeten Methoden wurden bis zu dieser Arbeit nur der knn-Vergleich sowie die Lineare Diskriminanzanalyse zur Klassifizierung von Proben in der Weinanalytik eingesetzt. Es wurden keine Literaturhinweise darauf gefunden, dass dies auch bei Neuronalen Netzwerken der Fall war.

Die Anzahl der beim Einsatz der verschiedenen Verfahren zur Gebietsvorhersage korrekt klassifizierten Proben ist in Tabelle 9 zusammengefasst. Da bei der Auswertung nicht jede Methode bei jedem Datensatz angewendet werden konnte, wurden die entsprechenden Felder leer gelassen. Es ist deutlich ersichtlich, dass die Vorhersagen mit Hilfe der alle Proben umfassenden Datensätze nicht sehr gut funktionierten und die besten Trefferquoten im Bereich von 55% lagen. Beim knn-Vergleich aller Proben kamen diese außerdem nur da-

Tab. 8: Ergebnisse der Herkunftsvorhersage mit Hilfe eines Neuronalen Netzwerks

Datensatz Nr.	Beschreibung	Korrekte Herkunftsvorhersage
3	Faktoren der Hauptkomponentenanalyse mit Hilfe der Korrelationsmatrix	55,93 % (33 von 59)
3a	Teilmenge Weinbaugebiete Donauland, Neusiedlersee und Weinviertel	77,27 % (34 von 44)
4	Faktoren der Hauptkomponentenanalyse mit Hilfe der Kovarianzmatrix	42,37 % (25 von 59)
4a	Teilmenge Weinbaugebiete Donauland, Neusiedlersee und Weinviertel	68,182 % (30 von 44)

durch zustande, dass dabei alle Proben ausschließlich den drei am häufigsten vorkommenden Weinbaugebieten zugeordnet wurden und so eine relativ hohe Wahrscheinlichkeit vorlag, die richtige Herkunft zu erraten. In dieser Beziehung verhielten sich sowohl die LDA als auch das Neuronale Netzwerk besser und zeigten keine derartig starke Bevorzugung der in größerer Anzahl vorhandenen Proben.

Bei der Auswertung mit Hilfe des knn-Vergleichs und des Neuronalen Netzwerks zeigte sich, dass die Verwendung der Daten aus der Hauptkomponentenanalyse mit Hilfe der Korrelationsmatrix (Datensatz 3) denen, die mit Hilfe der Kovarianzmatrix (Datensatz 4) gewonnen wurden, deutlich überlegen war. Diese Beobachtung wurde im Folgenden auch bei der Auswertung der Probenteilmenge aus den Gebieten Donauland, Neusiedlersee und Weinviertel (Datensätze 3a und 4a) bestätigt.

Die Trefferquoten bei der Auswertung der Daten dieser Teilmengen lagen durchwegs weit höher als bei den Datensätzen aus allen Proben. Die höchste Trefferquote wies hierbei der knn-Vergleich, allerdings in diesem Fall aus den bei der Hauptkomponentenanalyse gewonnenen Daten (Datensatz 3a), auf. Es ist jedoch sehr wichtig zu beachten, dass beim knn-Vergleich methodenbedingt nur zwischen den drei Weinbaugebieten der Probenteilmenge differenziert werden musste und damit die Wahrscheinlichkeit, die Herkunft zu erraten, bei 33,3% lag. Dasselbe galt auch für die LDA, die hier im direkten Vergleich mit der knn-Methode sogar besser abschnitt. Für die Auswertung mit Hilfe des Neuronalen Netzwerkes bestand diese Einschränkung nicht, es wurde in jedem Fall die Zugehörigkeit zu allen Weinbaugebieten getestet. Die dabei erzielten Trefferquoten, die auch im direkten Vergleich nicht viel schlechter ausfielen als die des knn-Vergleichs, müssen also in jedem Fall deutlich höher eingeschätzt werden.

Im Übrigen ergab der direkte Vergleich zwischen diesen beiden Methoden auch bei der Auswertung der Daten aller Proben beinahe keinen Unterschied zwischen den Trefferquoten. Das Neuronale Netzwerk lag dabei zum Teil sogar marginal vorne. Das kann als deutlicher Hinweis darauf gewertet werden, dass das Neuronale Netzwerk robuster auf die Probengesamtheit verzerrende Einflüsse reagiert. Da ein Neuronales Netzwerk im Gegensatz zu den beiden anderen eingesetzten Methoden versucht, die in den Datensätzen vorhandenen Zusammenhänge auf komplexe, nichtlineare Weise zu modellieren, ist es gut vorstellbar, dass es sich gerade deswegen gut dazu eignet, die nicht offensichtlichen Zusammenhänge zwischen von der Traube stammenden Präkursoren und den bei der Fermentation entstandenen Aromen zu modellieren. Da primäre Aromakomponenten, wie bereits oben erwähnt, keine besondere Rolle bei der gebietsspezifischen Aromatik eines Grüner Veltliner-Weines zu spielen scheinen, kam dieser Vorteil hier möglicherweise besonders zum Tragen.

Insgesamt scheinen Neuronale Netzwerke ein äußerst interessantes neues Verfahren für viele Fragestellungen der Weinanalytik zu sein. Dass sie bisher nicht umfangreich in dem Bereich eingesetzt wurden, liegt wohl auch zu einem nicht unerheblichen Teil an dem sehr hohen Berechnungsaufwand, den sie erfordern. Erst in den letzten Jahren wurde die dazu notwendige Computerleistung für die meisten Anwender verfügbar, bis dahin blieb der Einsatz von Neuronalen Netzwerken auf den Bereich von teuren Workstations oder sogar Großrechnern beschränkt. Aber selbst heute ist ihr Einsatz noch relativ zeitaufwändig. Für den Einsatz eines Neuronalen Netzwerks, so wie er hier beschrieben wurde, sind auch mit aktueller Computerhardware schnell Rechenzeiten im Bereich von einer Stunde und mehr notwendig, wobei das Training des Netzwerkes den größten Zeitaufwand benötigt.

Tab. 9: Vergleich der korrekt vorhergesagten Herkunftsgebiete bei allen eingesetzten Methoden (gerundet, n.a. = nicht ausgewertet)

Datensatz Nr.	Beschreibung	Korrekte Vorhersagen		
		knn-Vgl	LDA	Neur. Netw.
1	Konzentrationen aller freien Aromakomponenten	n.a.	40,68 %	n.a.
1a	Teilmenge Weinbaugebiete Donauland, Neusiedlersee und Weinviertel	n.a.	42,37 %	n.a.
2	Konzentrationen aller für die Herkunft signifikanten Aromakomponenten	52,63 %	n.a.	n.a.
2a	Teilmenge Weinbaugebiete Donauland, Neusiedlersee und Weinviertel	61,36 %	65,91 %	n.a.
3	Faktoren der Hauptkomponentenanalyse mit Hilfe der Korrelationsmatrix	54,24 %	n.a.	55,93 %
3a	Teilmenge Weinbaugebiete Donauland, Neusiedlersee und Weinviertel	81,82 %	n.a.	77,27 %
4	Faktoren der Hauptkomponentenanalyse mit Hilfe der Kovarianzmatrix	n.a.	n.a.	42,37 %
4a	Teilmenge Weinbaugebiete Donauland, Neusiedlersee und Weinviertel	n.a.	n.a.	68,18 %

Es konnte ebenfalls gezeigt werden, dass auch die beiden anderen getesteten Methoden bei richtiger Anwendung zur Vorhersage der Herkunftsgebiete geeignet sind. Der knn-Vergleich ergab die im Vergleich noch besseren Ergebnisse. Um diese zu erreichen, muss allerdings eine möglichst gleichmäßige Probenverteilung garantiert werden. Insgesamt erscheint eine Kombination mehrerer Methoden am geeignetsten. Da dafür identische Ausgangsdaten verwendet werden können, hält sich der dafür notwendige zusätzliche Aufwand in engen Grenzen und betrifft nur die tatsächlichen Auswertungsschritte. Mit zunehmender Leistung der verfügbaren Computer und der Software stellt dies zukünftig auch bei sehr großen Datenmengen kein wirkliches Hindernis dar.

Literatur

Arrhenius, S.P., McCloskey, L.P. and Sylvan, M. 1996: Chemical markers for aroma of Vitis vinifera var. Chardonnay regional wines. J. Agric. Food Chem. 44(4): 1085 - 1090

Amann, R. 2003: Schwarze Johannisbeeren und grüner Paprika in Cabernet. Schweiz. Z. Obst- und Weinbau 139(16): 6-9

Berente, B. (2004): HPLC-Analyse von Anthocyanen im Rotwein und Klassifizierung deutscher Rotweine mittels multivariater statistischer Methoden. - Diss. Friedrich-Schiller-Universität Jena, 2004

Clarke, R.J. and Bakker, J. (2004): Wine flavour chemistry. - Oxford: Blackwell, 2004

Henrion, R. und Henrion, G. (1995): Multivariate Datenanalyse : Methodik und Anwendung in der Chemie und verwandten Gebieten. - Berlin: Springer, 1995

Majdak, A., Herjavec, S., Orlić, S., Redžepović, S. and Mirosević, N. 2002: Comparison of wine aroma compounds produced by Saccaromyces paradoxus and Saccaromyces cerevisiae strains. Food Technol. Biotechnol. 40(2): 103-109

King, A.J. and Dickinson, R.J. 2003: Biotransformation of hop aroma terpenoids by ale and lager yeasts. FEMS Yeast Research 3(1): 53-62

Klimmek, A.(2003): Bestimmung des geografischen Ursprungs von Weinen mittels Multikomponentenanalyse und multivariater Statistik. - Diss. TU Berlin, 2003

ÖWM (2005): Dokumentation Österreichischer Wein. - Wien: Österreichische Weinmarketingservice GmbH, 2005

Papargyriou, E. (2003): Veränderung von glykosidisch gebundenen Sekundärmetaboliten bei Vitis vinifera L. (cvs. Gewürztraminer und Riesling) in Zusammenhang mit Traubenreife, Weinbereitung und Weinlagerung. - Diss. Justus-Liebig-Universität Giessen, 2003

Rapp, A. 1998: Volatile flavor of wine : Correlation between instrumental analysis and sensory perception. Nahrung 42: 351-363

Rapp, A., Güntert, M. und Heimann, W. 1985: Beitrag zur Sortencharakterisierung der Rebsorte Weißer Riesling. Z. Lebensmittel-Unters. -Forschung 181: 357 - 361

Rapp, A. und Hastrich, H. 1978: Gaschromatographische Untersuchungen über die Aromastoffe von Weinbeeren, Teil 3: Die Bedeutung des Standortes für die Aromastoffzusammensetzung der Rebsorte Riesling. Vitis 17: 288-298

Rapp, A. and Mandary, H. 1986: Wine aroma. Experientia 42: 873-884

Seeber, R., Sferiazzo, G., Leardi, R., Dalla Serra, A. and Versini, G. 1991: Multivariate data analysis in classification of musts and wines of the same variety according to vintage year. J. Agric. Food Chem. 39: 1764 - 1789

Versini, G., Inama, S. and Sartori, G. 1981: A capillary column gaschromatographic research into the terpene constituents of Riesling Renano (Rhine Riesling) wine from Trentino Alto Adige : Their distribution within berries, their passage into must and their presence in the wine according to different wine making procedures. Organoleptic considerations. Vini d'Italia 23: 189-211

Wallner, E., Kreuz, S., Flak, W. und Nikiforov, A. 1999: Charakterisierung des österreichischen Rieslings mittels GC-MS und multivariater Datenanalyse. Mitt. Klosterneuburg 49(1): 14-22

Williams, P.J., Strauss, C.R. and Wilson, B. 1981: Classification of the monoterpenoid composition of Muscat grapes. Am. J. Enol. Vitic. 32: 230-235

Wondra, M. and Berovi, M. 2001: Analyses of aroma components of Chardonnay wine fermented by different yeast strains. Food Technol. Biotechnol. 39: 141-148

Zirilli, J. (1997): Financial prediction using neural networks. - London: Int. Thomson Computer Press, 1997

Manuskript eingelangt am 28. August 2006