Apidologie, 1971, 2 (1), 49-65.

ÉTUDE D'UNE MÉTHODE BIOMÉTRIQUE ET STATISTIQUE PERMETTANT LA DISCRIMINATION ET LA CLASSIFICATION DE POPULATIONS D'ABEILLES (*Apis mellifica* L.)

Eine biometrische und statistische Methode zur Untersuchung und Klassifizierung von Bienenvölkern (Apis mellifica L.)

Richard TOMASSONE et Jean FRESNAYE

Laboratoire de Biométrie,
Centre national de Recherches zootechniques, I.N.R.A.,
78 — Jouy-en-Josas
Station expérimentale d'Apiculture, Centre de Recherches d'Avignon, I.N.R.A.,
84 — Montfavet

SUMMARY

STUDY OF A BIOMETRICAL AND STATISTICAL METHOD TO DISCRIMINATE AND SEPARATE BEES' POPULATIONS

This paper is a case study of the application of some multivariate methods to discriminate and classify bees' progenies. The study is based on fifty progenies with eight variables computed on them.

First we use canonical variates analysis — or dispersion analysis — which permits us to affirm that the observed differences between progenies are significant, and that these differences are due to specific factors. But this analysis would not be sufficient if it was not possible to make clusters of progenies in order to classify them in a hierarchical way. A measure of distance between two populations is introduced; it permits to use two numerical taxonomy methods and to compare their respective proprieties. But with the help of this distance it is also possible to allocate a next progeny to one population or the other according to its « nearness ».

All the computer's programs are available; they are written in Fortran IV.

RÉSUMÉ

Cette étude expose l'application de quelques méthodes d'analyses multivariables pour discriminer et classer des populations d'abeilles. L'étude est faite à l'aide de 50 souches sur lesquelles ont été calculées huit variables.

Nous utilisons d'abord l'analyse des variables canoniques — ou analyse de dispersion — qui nous permet d'affirmer que les différences observées entre les souches sont significatives et qu'elles sont dues à des facteurs spécifiques. Cette analyse serait insuffisante s'il n'était pas possible de former des groupes de souches pour les classer de manière hiérarchique. Nous introduisons une mesure de la distance entre deux propositions qui nous permet d'utiliser deux méthodes de taxinomie numérique et de les comparer en présentant leurs propriétés. Mais à l'aide de cette distance il est également possible de rattacher une nouvelle souche à l'une ou l'autre des populations en fonction de sa proximité.

Tous ces programmes d'ordinateur sont disponibles; ils sont écrits en Fortran IV.

INTRODUCTION

L'étude qui va être présentée est une illustration de l'utilisation des méthodes statistiques à plusieurs variables pour discriminer des populations puis pour les classer selon des critères automatiques, donc objectifs (dans un sens qui reste à préciser, d'ailleurs). Les populations représentent ici des souches d'abeilles et les variables qui permettent de les décrire sont des mesures de caractéristiques biométriques.

MATÉRIEL BIOLOGIQUE ET MÉTHODES DE MESURES

De nombreux caractères morphologiques ont été étudiés en biométrie de l'abeille, notamment par Goetze (1926 à 1963). Tous ne présentent pas le même intérêt ni la même facilité d'utilisation. Il est certes utile d'analyser suffisamment de critères discriminatoires, mais il est également indispensable que le temps requis par les analyses reste dans les limites des possibilités matérielles des utilisateurs. C'est pourquoi, à la suite de Ruttner et Mackensen (1954-1963), nous nous étions limités à 5 caractères dans une première étude sur l'abeille noire française (Fresnaye, 1965) : la couleur, l'index cubital, la pilosité du 5eme tergite abdominal, la largeur du tomentum sur le 4eme tergite, la longueur de la langue.

La coloration de l'exosquelette, détermination de la présence et de l'importance de taches jaunes sur les premiers tergites de l'abdomen, a constitué à l'origine la base de la taxinomie des races d'abeilles. Elle est actuellement très controversée, surtout en raison de sa grande variation naturelle (Goetze 1940, Ruttner 1952-1968, Giavarini 1953). Nous l'avons écartée provisoirement, nous réservant de l'étudier séparément et de l'adjoindre ultérieurement à nos analyses.

Les méthodes de mesure de ces caractères morphologiques sont décrites dans les publications citées ci-dessus. Seul l'index cubital, dont la méthode de mesure reste la même que précédemment, est désormais utilisé dans une méthode statistique différente. Le rapport de deux variables aléatoires est une nouvelle variable aléatoire, ici l'index cubital, dont l'étude est en général complexe même si l'on connait les deux lois de distribution des variables qui servent à le calculer. Pour cette raison, les statisticiens n'aiment en général pas l'utiliser directement; ils préfèrent le faire apparaître plus « logiquement »; par exemple, en calculant la différence du logarithme des deux variables, c'est-à-dire le logarithme du rapport. En outre, le simple fait de « condenser » deux variables en une seule entraîne une perte d'information qui pourrait être intéressante pour une étude de discrimination. Ici, nous avons utilisé une méthode mise au point par Myint Tin (1965), pour calculer valeur moyenne et variance sur un échantillon.

Nous nous référerons par la suite aux huit variables suivantes :

x_1 = Index cubital
x_2 = Longueur des poils
x_3 = Tomentum
x_4 = Longueur de la langue
x_5 à x_8 les variances des quantités précédentes.

Les souches d'abeilles que nous avons utilisées pour notre étude sont des groupes de colonies dont le degré de parenté est variable suivant leur origine. On distingue :

a) les souches constituées à Montfavet à partir de reines sœurs fécondées dans les mêmes conditions. Les fécondations naturelles ont lieu en même temps dans le même rucher de fécondation. Les inséminations artificielles sont pratiquées avec le sperme de mâles d'origine connue.

b) les souches des écotypes Bretagne, Essonne et Landes, constituées dans les régions d'origine, avec des colonies n'ayant pas subi de croisement avec des races ou des écotypes étrangers mais sans consanguinité.

c) les souches de races étrangères pour lesquelles la pureté de la race est la seule certitude.

Cette étude est réalisée à partir de 50 souches d'abeilles représentant 424 colonies. L'index cubital a été calculé sur des échantillons de 30 à 100 abeilles; la longueur des poils, le tomentum et la longueur de la langue mesurés sur 10 abeilles par colonie. Ces analyses biométriques représentent un total d'environ 73 000 mesures.

ANALYSE DES DONNÉES
LES STATISTIQUES ÉLÉMENTAIRES

1. — *Généralités*

Nous avons de très nombreuses fois, dans d'autres publications, distingué les deux types d'approche d'analyse statistique d'un problème. D'un côté des méthodes rigoureuses s'appuyant sur des conditions bien définies permettent d'effectuer des tests, c'est-à-dire permettent d'attacher une probabilité à un résultat : c'est évidemment le stade le plus élaboré, celui qui permet de conclure avec une certaine rigueur. D'un autre côté, des méthodes d'investigation, c'est-à-dire des méthodes qui constituent plus une source d'hypothèses à vérifier qu'une fin en elles-mêmes; ces hypothèses peuvent être ensuite soumises aux tests rigoureux dont nous venons de parler. Sans vouloir négliger les premières méthodes, nous nous attacherons à montrer ici les possibilités des secondes.

Les études biologiques sont multivariables par nature; en effet, comment décrire un matériel fluctuant sans attacher plusieurs critères à sa description ? Si nous disons que l'objet que nous analysons vole (c'est déjà une description !) nous avons déjà fait un tri mais bien insuffisant car nous pouvons encore inclure dans notre analyse des oiseaux, des insectes et pourquoi pas des avions ? Si nous précisons que les objets ont des ailes et qu'ils peuvent produire du miel nous pouvons nous limiter aux apides mais là encore toute étude un peu approfondie nous montre qu'il existe des groupes différents plus ou moins voisins. Certains groupes se distinguent si aisément à l'aide d'un seul critère, qu'il paraît inutile (d'un simple point de vue taxinomique s'entend, puisque des considérations économiques pourraient remettre en question la distinction) de faire appel à plusieurs. Par contre, pour d'autres il est manifestement impos-

sible d'établir des coupures, des classes et même de savoir quels sont les meilleurs critères à prendre pour mieux les distinguer. C'est à ce moment que les techniques à plusieurs variables peuvent entrer en action. Bien sûr, il existe des démonstrations rigoureuses pour ces techniques; elles sont généralement complexes, au moins pour le biologiste toujours effrayé par un appareil mathématique trop inconnu. Nous nous attacherons donc ici à utiliser des méthodes dans le but unique de *voir*, *d'ordonner* et de *classer* les éléments, ce que le simple examen des données de base ne nous permet pas toujours de faire.

2. — *Les données de base*

Ce sont les huit variables mesurées sur chaque élément, ici c'est une colonie d'abeilles; (en réalité ces variables sont des moyennes et des variances calculées sur des échantillons de trente à cent abeilles d'une ruche). Chaque colonie appartient à une souche et ce sont ces dernières qu'il faut discriminer puis classer. Ces variables forment ce que nous appellerons la matrice des données de base; c'est un tableau à huit colonnes dont le nombre total de lignes est égal au nombre de colonies soit 424. Mais ici il nous faut une colonne supplémentaire pour bien caractériser une ligne de la matrice; il faut que nous sachions que cette ligne (plus précisément la colonie représentée par cette ligne) appartient à une souche déterminée; pour cela nous dirons que la matrice des données de base est partitionnée et nous entendons par là que les colonies appartenant à une même souche ont été mises côte à côte dans la matrice. Si les souches avaient été représentées par leurs huit moyennes, la matrice n'aurait pas été partitionnée. Et même avec les 424 lignes la matrice aurait pu ne pas être partitionnée si nous avions voulu discriminer et classer les colonies sans nous préoccuper de leur appartenance a priori à une souche bien déterminée. Le fait que l'étude porte sur les souches et que nous possédions ces deux niveaux de variation va nous permettre de faire une analyse beaucoup plus fine.

Nous pouvons décomposer la variation globale des 424 observations en deux parties, l'une due à la variation entre les 50 souches, l'autre due à la variation à l'intérieur des souches; c'est le schéma classique de l'analyse de variance à un facteur contrôlé : la somme des carrés des écarts à la moyenne de toutes les observations est égale à la somme de deux sommes de carrés : l'une mesure les écarts entre les moyennes des souches et la moyenne générale, l'autre englobe la somme de tous les écarts des observations à leur propre moyenne de groupe; nous avons l'égalité classique de l'analyse de variance

$$T = B + W$$

où
- T représente la somme des carrés totale,
- B (comme between) celle entre souches,
- W (comme within) celle intérieure aux souches.

Chaque somme est affectée de degrés de liberté tels que l'égalité valable sur les sommes l'est aussi sur les degrés.

Si nous faisons intervenir plusieurs caractères cette égalité est toujours vérifiée ; mais il s'agit maintenant d'égalité sur des tableaux, des matrices. Ces matrices carrées sont constituées des sommes de carrés et de produits des variables analysées. Ici les matrices sont de dimension 8 puisque nous avons huit variables ; elles sont symétriques et nous pouvons ne considérer qu'une moitié du tableau. Sur la diagonale sont les sommes de carrés et une somme est repérée par son numéro de ligne, ou par son numéro de colonne puisque les deux sont identiques. A un numéro de ligne i et un numéro de colonne j correspond la somme des produits relative à la variable i et à la variable j.

A l'aide de ces tableaux il est possible de calculer deux nouvelles matrices, celles des coefficients de corrélation entre les variables ; nous les noterons respectivement R_b et R_w. Il faut bien noter que ces corrélations ont des significations bien différentes : l'une traduit des corrélations entre les moyennes des souches, l'autre des corrélations internes aux souches ; bien sûr, ces dernières peuvent varier d'une souche à l'autre, mais nous ferons l'hypothèse qu'il est possible d'utiliser une valeur moyenne calculée à partir de W. Dans la mesure où nous utilisons moyennes et variances nous sommes déjà en contradiction avec cette hypothèse et il faut toujours s'en souvenir au cours de l'analyse. Cette différence entre les deux niveaux est illustrée par la figure 1 ; nous avons

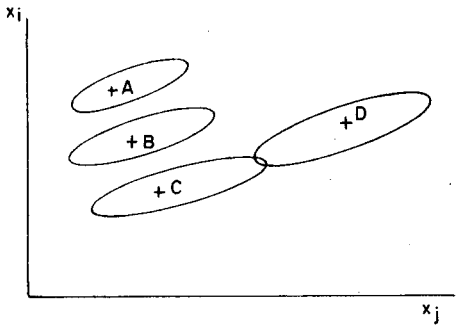

FIG. 1. — *Corrélation interne à chaque population symbolisée par un nuage elliptique.*
ABB. 1. — *Korrelation innerhalb jeder Population, dargestellt durch eine elliptische Wolke*

symbolisé la corrélation interne à chaque population par un nuage elliptique (qui nous rappelle que les nuages de point elliptiques sont ceux trouvés dans le cas de distributions normales).

Nous comprenons intuitivement que les différences entre les valeurs des coefficients de corrélation vont être d'une grande importance pour discriminer et classer les souches : des valeurs de même sens, cas de C et D, ne permettront pas de les séparer alors que des valeurs opposées le permettront beaucoup mieux.

3. — *Les données de travail synthétiques*

Ce sont les matrices R_b et R_w la première à 49 degrés de liberté, la seconde à 374 degrés de liberté. (Les valeurs des coefficients ont été multipliées par 100).

Matrice R_b ($r_{.01} = 35$, est la valeur significative au seuil 1 % indiquée par * à droite du coefficient).

100									
— 78*	100								
63*	— 65*	100							
61*	— 49*	52*	100						
75*	— 53*	61*	54*	100					
02	— 32	— 05	— 17	— 01	100				
17	— 27	16	08	— 01	44*	100			
— 23	10	— 38	— 28	— 31	18	24	100		

Matrice R_w ($r_{.01} = 13$)

100									
01	100								
— 02	— 14*	100							
04	14*	— 01	100						
34*	— 04*	— 03	00	100					
04	— 12	— 06	06	03	100				
05	— 06	— 05	— 08	05	— 03	100			
00	— 07	01	— 04	00	04	02	100		

Nous voyons que les 5 premières variables sont correlées dans R_b alors que les valeurs dans R_w sont généralement très faibles.

ANALYSE APPROFONDIE

1. — *Discrimination des souches*

Dans un premier temps nous cherchons à savoir si les souches appartiennent à une population unique ou si au contraire elles diffèrent; nous ne cherchons pas à savoir pour l'instant comment les regrouper et nous nous contentons d'un test global : existe-t-il ou non des différences entre les souches ? L'analyse séparée de chaque variable a permis de dire que pour les sept premiers critères il existait des différences entre les souches. Dans la mesure ou les variables sont correlées il est permis de se demander si nous ne refaisons pas les mêmes analyses quand nous regardons les variables l'une après l'autre, chaque variable représentant alors une « variation sur un même thème ». En plus chaque variable est significative, mais n'est-il pas possible d'en trouver une

nouvelle encore plus significative qui puisse synthétiser de façon valable la variabilité ? Pour cela nous faisons une transformation de variables qui doit fournir la plus grande différence de la variation entre souches par rapport à la variation interne des souches. Cette transformation de variables est linéaire c'est-à-dire que nous nous contentons d'une combinaison linéaire des variables de départ ; c'est encore une hypothèse restrictive.

En fait on trouve toutes les combinaisons linéaires possible en une seule fois en calculant les valeurs propres et les vecteurs propres de la matrice B associée à W, c'est-à-dire les valeurs et les vecteurs V_i tels que $BV_i = \lambda_i WV_i$. Les nouvelles variables obtenues sont ce que nous appellerons les *variables canoniques* ;

TABLEAU 1. — *Valeur des coefficients des variables canoniques, test des coefficients*

TABELLE 1. — *Koeffizientenwerte der kanonischen Variablen ; Koeffiziententest*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	*dl*	χ^2
y_1	*44*	— *54*	*36*	*49*	*38*	— 09	01	07	392	2255.71**
y_2	— *37*	*49*	18	— 04	*76*	— 04	— 10	— 04	336	1371.19**
y_3	— 16	21	— 17	*92*	— 21	— 23	04	03	282	958.35**
y_4	— 33	11	*87*	01	— 25	07	20	— 11	230	642.18**
y_5	— *73*	— *46*	— 22	19	*34*	*38*	18	08	180	385.90**
y_6	— 24	— 39	— 05	— 03	01	— 38	— *86*	— 20	132	214.33**

leur nombre est égal à la plus petite des deux valeurs : nombre de variables ou nombre de populations diminué d'une unité. Ainsi avec huit variables et cinquante populations nous avons huit variables canoniques ; si, toujours avec huit variables, vous avions eu six populations il n'y aurait eu que cinq variables canoniques. Ce fait rejoint le cas particulier de la fonction discriminante de Fisher à deux populations qui est unique quel que soit le nombre de variables utilisées dans l'étude.

Il est possible de rechercher si toutes ces variables sont significatives ; ce faisant on peut étudier la variabilité dans un espace de dimension inférieure. Comme les nouvelles variables sont indépendantes et classées par ordre de discrimination décroissante il est possible de voir les critères dont l'apport est le plus important et de les classer selon le pouvoir discriminant. Ainsi nous avons trouvé six variables canoniques significatives au seuil 1 % ; la diminution est faible, mais dans la mesure où les variables canoniques sont des combinaisons de variables il est possible d'analyser cette variabilité en des composantes indépendantes dont la signification biologique n'est pas sans intérêt. Dans le tableau 1 le test est fait à l'aide d'un test du χ^2 (Rao, 1965) ; la signification des variables est déterminée en regardant l'ordre de grandeur respectif des différents coefficients dans la combinaison linéaire.

La première variable associe avec un poids sensiblement égal les variables 1, 3, 4, 5 qu'elle oppose à la variable 2 ; l'opposition est traduite par le signe du coefficient de $\times 2$. Elle peut faire penser à un rapport entre les deux groupes de

variables. Les grandes valeurs de cette variable s'appliquent aux éléments pour lesquels l'index cubital, sa variabilité, le tomentum et la longueur de la langue sont élevés alors que la longueur des poils est faible ; les petites valeurs correspondent au cas opposé. La seconde variable est surtout liée à 5 (variabilité de l'index cubital), mais elle lui associe à un moindre degré 2-1. La présence d'une variable de la forme 5-1 doit suggérer que le coefficient de variation de l'index cubital peut être un discriminateur intéressant. Les variables canoniques 3 et 4 ne font intervenir chacune qu'une seule variable : la longueur de la langue pour la troisième et le tomentum pour la quatrième ; elles sont donc simples à utiliser ; il est bien entendu que c'est la part plus originale de ces variables qui intervient ici puisque nous les avons toutes deux déjà rencontrées dans la première. La variable 5 est une sorte de complément à la seconde : l'index cubital à la part la plus forte, mais il s'oppose à sa variabilité, comme la longueur des poils ; elle pourrait donc nous suggérer l'analyse des coefficients de variation de ces deux variables. Enfin, la dernière variable est presque entièrement liée à la variabilité du tomentum.

Ainsi nous utilisons ces résultats pour avoir une première image de la variabilité des souches que nous étudions, mais surtout nous voyons à travers ces résultats une orientation nouvelle dans l'analyse des données élémentaires.

Quel est le sens de ces variables biologiques forgées à partir de critères naturellement utilisés ? Ont-elles un sens ? La relation de ces variables biologiques avec les études biométriques antérieures apparaît plus ou moins clairement mais il faut considérer l'éventail d'échantillons soumis aux tests. Il s'agit d'une importante majorité de colonies d'abeilles de race noire, d'écotype provençal. Cet écotype est confronté à de petits groupes d'échantillons d'écotypes d'abeilles des régions parisienne, bretonne, landaise, d'une part et de quelques échantillons d'abeille de races étrangères, *Apis mellifica ligustica* et *Apis mellifica intermissa*. On sait déjà que chez l'abeille noire l'index cubital est faible (1,70 — 1,80) et la longueur des poils importante (0,46 — 0,50). Chez les races géographiquement les plus voisines de la race noire on trouve pour *ligustica* un index cubital de 2,30 et des poils de 0,30 et chez *carnica* un index cubital de 2,70 et des poils de 0,30 également. Cette liaison dans une race entre index cubital faible et poils longs ou au contraire index cubital élevé et poils courts est particulièrement soulignée par la première variable canonique de notre étude. De même cette variable relie le tomentum faible et la langue courte à l'index cubital faible d'une part ; le tomentum large et la langue plus longue, à l'index cubital élevé d'autre part. Cette relation avec les données taxinomiques acquises est également visible pour les races concernées par notre étude.

2. — *Classification des souches*

1. *Principes.*

Nous savons maintenant qu'il existe des différences entre les souches et

nous avons même une idée sur la dimension de la variabilité, sur les critères les meilleurs pour l'analyser. Nous pouvons aller plus loin et essayer de regrouper les souches en fonction de leur ressemblance. Nous faisons alors appel aux méthodes de la taxinomie numérique ; nous allons voir que les critères statistiques n'interviennent pas dans cette analyse. Il existe un très grand nombre de techniques pour classer des éléments et il est délicat de voir les grands principes qui les dirigent ; pour une présentation générale cf. Sokal et Sneath, 1963, ou pour une présentation plus succincte : Millier et Tomassone, 1969.

2. *Choix d'une mesure de la ressemblance.*

Il faut avant toute étude de classification définir une distance, c'est-à-dire une mesure de la ressemblance entre les éléments qu'on se propose de classer. A l'heure actuelle presque toutes les techniques employées demandent la connaissance de tous les couples de distances entre éléments : lorsqu'il y a N éléments, il y a N (N-1) /2 distances ; ce nombre qui croît rapidement avec le nombre d'éléments constitue une limite pratique importante dans les études de classification (1).

Le choix du type de la distance est naturellement essentiel puisque c'est avec ces mesures que nous allons regrouper des éléments, fusionner des groupes. Dans notre étude, comme nous possédons une estimation de la variabilité résiduelle grâce à la matrice W nous choisissons la distance de Mahalanobis ; celle-ci présente de très grands avantages puisqu'elle a un sens statistique : la distance entre C et B, (fig. 1) étant plus grande que celle entre C et D bien que géométriquement la dernière soit plus grande. De plus elle est totalement indépendante de l'échelle des variables de départ, ce qui n'est pas le cas d'autres mesures de distances.

3. *Les différentes méthodes de classification.*

Si nous ne considérons que les méthodes n'admettant aucun recouvrement entre les groupes créés nous avons deux grands types de méthodes.

a) *Les méthodes divisives :* dans lesquelles la première opération consiste à classer les éléments en deux groupes aussi distincts que possible : puis, sur chacun des groupes créés, on recommence le processus jusqu'à ce qu'on décide de s'arrêter, par exemple lorsqu'une nouvelle coupure n'apporte plus aucune amélioration significative. Les phyto-sociologues utilisent fréquemment l'analyse d'association de Williams et Lambert qui appartient à ce type.

b) *les méthodes agglomératives* dans lesquelles on commence par regrouper les éléments les plus voisins ; puis on rattache progressivement les autres à des niveaux plus ou moins éloignés. Ainsi, on aboutit à une sorte d'arbre généa-

(1.) Une centaine d'éléments sur les petits ordinateurs, du type 1130 IBM 8K, constitue une limite supérieure.

logique (1) — *un dendrogramme* — qui synthétise tout ce qu'il est nécessaire de connaître sur la classification. Sneath, 1968, considère que dans les principes qui régissent la formation des groupes existent trois *concepts* distincts :

— un concept de *masse* : les éléments entrent dans le groupe avec un poids unitaire, et viennent se fondre dans lui pour en augmenter le poids.

— un concept de *densité* : les centres des groupes contiennent plus d'éléments que les limites sur lesquelles on ne trouve que très peu d'éléments, ceux qu'il n'est guère possible de classer.

— un concept de *réseau* : les liens entre les éléments sont les facteurs les plus importants : chaque élément conserve son individualité au sein du groupe déjà créé au lieu de se perdre dans celui-ci. On attache plus d'importance aux relations entre les distances qu'aux valeurs exactes des distances elles-mêmes ; cet aspect est très important puisque le choix de ces dernières est souvent très arbitraire. On peut, en effet, raisonnablement admettre que des formes mathématiques différentes des mesures des distances entraînent de fortes variations des valeurs absolues ; par contre, les inégalités qu'elles vérifient seront souvent respectées. Les méthodes utilisant le concept de réseau attachant plus d'importance aux relations entre les distances qu'aux valeurs exactes de ces distances ont, à ce titre, un très grand intérêt. Malheureusement des effets de chaîne peuvent s'introduire : il est quelquefois possible d'en limiter les conséquences (Wishart, 1968).

Si nous revenons sur les deux premiers concepts, nous pouvons dire que les méthodes qui s'en inspirent cherchent à rendre optimum l'homogénéité à l'intérieur des groupes ; plus précisément on peut distinguer quatre étapes dans le processus de formation des groupes :

a) comment créer les premiers groupes ? Par exemple en prenant les éléments les plus voisins.

b) comment fusionner des éléments nouveaux à des groupes déjà créés ? Par exemple en imposant à la distance interne du groupe d'être inférieure à une limite fixée.

c) comment déterminer qu'une nouvelle fusion est inutile ? par exemple si la distance de l'élément au groupe est supérieure à une valeur fixée.

d) comment affecter à d'autres groupes des éléments déjà classés de façon à se prémunir contre certains effets néfastes — par exemple des effets de chaîne ou de migration — ; cette nouvelle affectation se fait à l'aide d'échanges entre les groupes.

4. *Les méthodes choisies.*

Notre choix s'est porté sur deux méthodes faisant intervenir des concepts différents :

(1.) C'est une ressemblance formelle, car le fait que deux branches d'un arbre soient voisines n'implique aucune relation phylétique mais de simples relations phénétiques.

— celle de Roux M., 1968 qui par l'allure assez enchaînée du dendrogramme auquel elle aboutit, illustre bien le concept de réseau (fig. 2).

— celle de Van den Driessche, 1965 qui fournit un dendrogramme d'allure beaucoup plus lourde (fig. 3).

Mais ici les groupes des souches sont suffisamment bien tranchés pour que l'algorithme de calcul n'ait que très peu d'influence sur le résultat. Comme nous voyons apparaître dans les deux cas les mêmes groupes, nous pouvons avoir une assez grande confiance dans leur validité.

3. — *Interprétation des résultats*

Le tableau 2 (p. 64) indique pour chaque souche étudiée dans l'ordre de classement établi par le dendrogramme de la figure 2, l'origine, l'écotype ou la race, l'année de création, ainsi que les divisions principales déterminées par le den- drogramme.

On constate que :

a) L'écotype Provence, entièrement localisé dans la partie gauche du dendrogramme, forme un groupe bien différencié des autres écotypes. La distance maximale de liaison de ce groupe est de 4,77. Nous verrons en outre plus loin que des sous-groupes nettement caractérisés apparaissent dans cet écotype.

b) A la suite de l'écotype Provence, on trouve les écotypes Bretagne et Essonne (région parisienne). Le nombre des échantillons dans ces deux écotypes est assez faible. Il semble cependant dès à présent que la morphologie de ces deux groupes est assez similaire. La distance généralisée qui les sépare n'est que de 2,27 et leur branche se rattache au dendrogramme à une distance D^2 de 6,82.

c) L'écotype *Landes* vient ensuite; avec un D^2 de 10,59 il est également nettement différencié des écotypes précédents.

d) Viennent ensuite les différents groupes d'abeilles hybrides de *Apis mellifica mellifica* x *Apis mellifica ligustica* reliés entre eux aux niveaux de 3,40 et 5,91; ils se rattachent au dendrogramme par des D^2 de 12,19 et 13,40.

e) Les groupes de colonies *Apis mellifica ligustica* de race pure se raccordent beaucoup plus bas au dendrogramme avec des D^2 de 35,12 et 38,18.

f) Les quelques échantillons d'*Apis mellifica intermissa* sont, eux aussi, parfaitement différenciés; leur D^2 est de 58,34.

g) Des ramifications apparaissent à l'intérieur de l'écotype *Provence*. Il semble que la finesse de discrimination de la méthode nous permette d'isoler des sous groupes dont les causes de variations morphologiques sont encore mal déterminées. On distingue notamment un sous groupe d'analyses postérieur à 1965 et un sous groupe d'analyses antérieur à cette date. Ce dernier est lui-

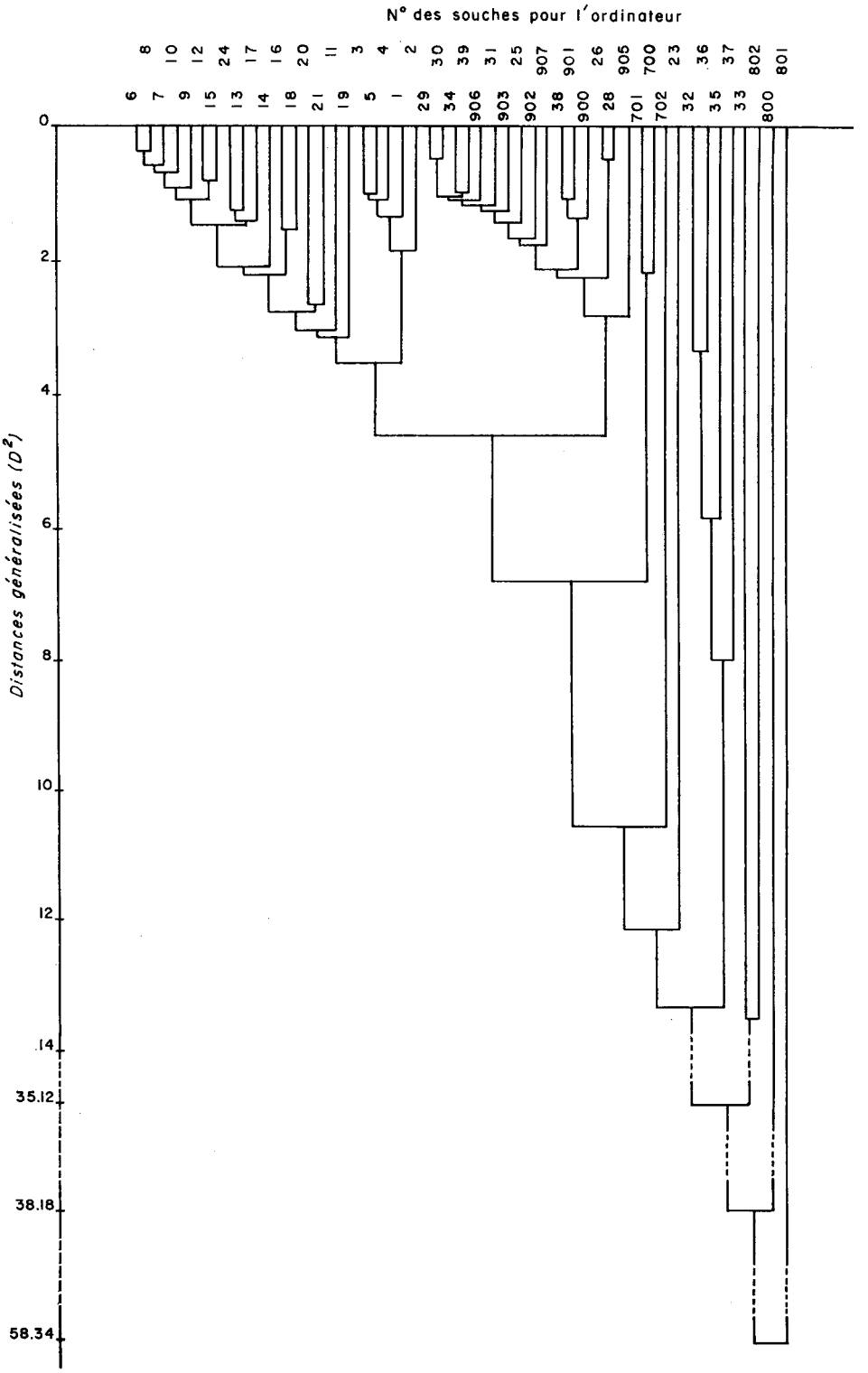

FIG. 2. — *Dendrogramme de classification des souches par la méthode de Roux.*
ABB. 2. — *Dendrogramm zur Klassifizierung der Stämme nach der Roux-Methode*

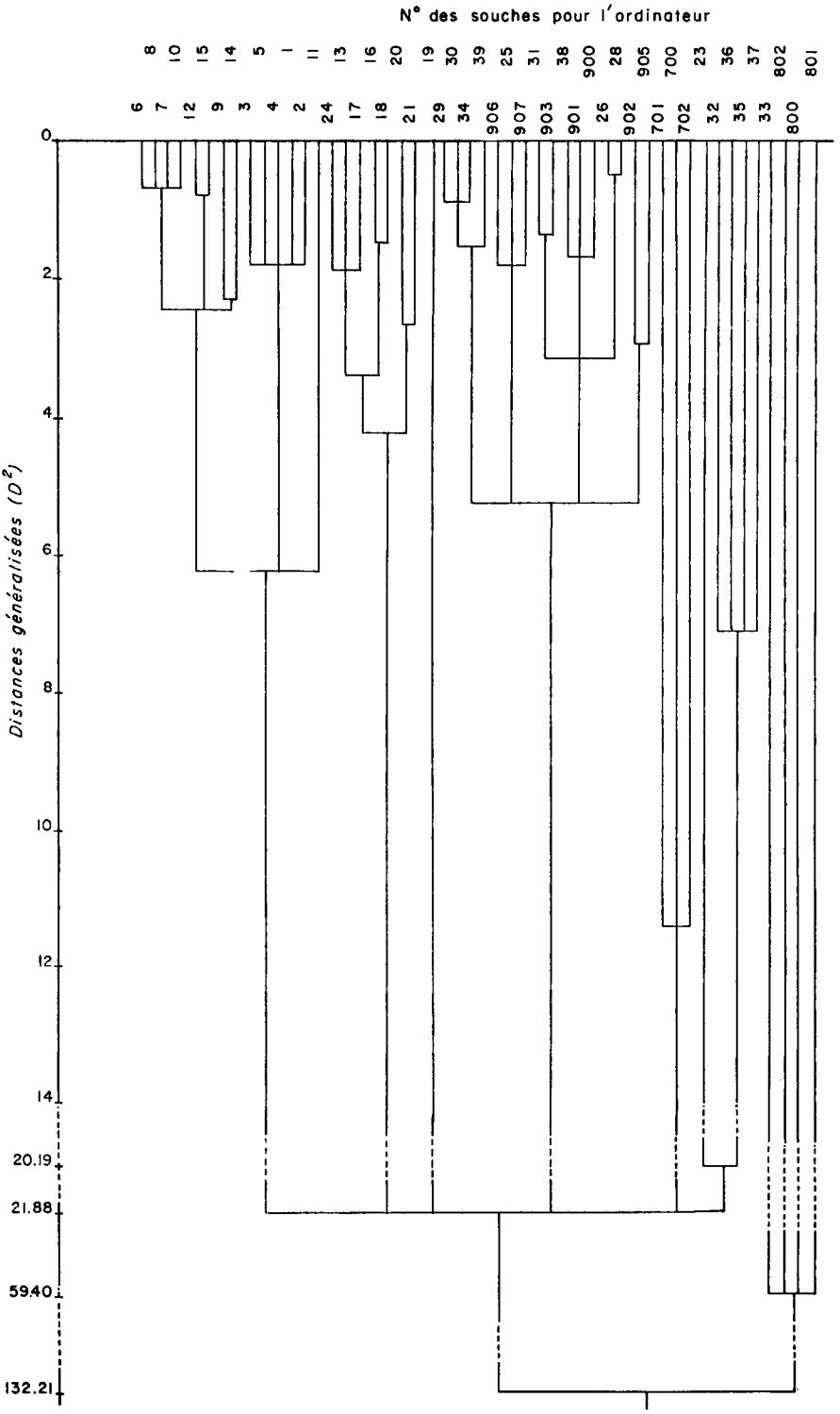

FIG. 3. — *Dendrogramme de classification des souches par la méthode de Van den Driessche.*

ABB. 3. — *Dendrogramm zur Klassifizierung der Stämme nach van den Driessche.*

même divisé en années 1960 et 1961 situées de part et d'autre des années 1962-63-64 qui sont mélangées. De nouvelles expériences et une étude plus poussée devraient nous permettre de définir les origines de cette classification au niveau des groupes annuels d'analyses biométriques.

4. — *Classement d'une nouvelle souche*

Nous avons eu affaire jusqu'à maintenant à une étude statique, dans la mesure où un nombre bien défini de souches était seul analysé. Nous allons maintenant imaginer que nous ayons une souche nouvelle avec les valeurs des 8 variables. Comment l'intégrer dans la classification précédente ? Le plus simple consiste à calculer sa distance de Mahalonobis avec toutes les souches existantes ; nous choisirons la plus petite de ces distances et nous pourrons la rattacher à la souche dont elle est la plus voisine. Nous verrons ainsi si elle se rattache à un niveau élevé dans le dendrogramme : dans ce cas, elle est très proche d'une souche déjà analysée ; dans le cas opposé où elle se rattache très bas, on peut affirmer qu'elle est différente des souches de base, (Cacoullos, 1965). Nous pouvons alors envisager de « créer » une nouvelle souche dans la classification.

CONCLUSION

Nous avons exposé les principes et les possibilités d'une méthode biométrique et statistique nouvelle permettant la discrimination et le classement des différentes races et écotypes d'abeilles (*Apis mellifica* L.). D'un point de vue pratique l'ensemble de tous les calculs est « programmé ». C'est-à-dire que l'analyse est peut-être complexe mais une fois comprise, la part matérielle des calculs est négligeable puisqu'un petit ordinateur, comme un IBM 1130 8K peut s'en charger en une heure, sortie graphique comprise.

Les bases taxinomiques et leur classement dans les dendrogrammes ne peuvent être considérés comme établis que pour l'écotype Provence de la race Noire française (*Apis mellifica mellifica*). Nous nous proposons d'établir ces bases pour les autres races d'abeilles domestiques et, dans la mesure du possible, de leurs écotypes. Nous disposerons alors d'une méthode dont la précision est inconnue actuellement dans les analyses biométriques chez l'abeille.

Reçu pour publication en septembre 1970.
Eingegangen im September 1970.

ZUSAMMENFASSUNG

Die vorliegende Arbeit erläutert die Anwendung statistischer Methoden mit mehreren Variablen zur Unterscheidung und nachfolgenden Klassifizierung von Bienenpopulationen nach automatischen Kriterien, wobei es sich um Bienenstämme handelt. Die Variablen, die

ihrer Beschreibung zugrunde liegen, sind folgende biometrische Daten : Cubitalindex, Haarlänge, Filzbinde und Rüsselänge.

Die von uns vorgeschlagenen Techniken laufen hauptsächlich darauf hinaus, sehr grosse Tabellen von Daten zu bearbeiten, die zum Betrachten, Ordnen und Klassifizieren dienen sollen, was mit der einfachen Prüfung der Grundeigenschaften nicht immer möglich ist. Zu Beginn muss überprüft werden, ob und mit welchen Eigenschaften eine Unterscheidung möglich ist. Das ist Aufgabe der Analyse der kanonischen Variablen. Sie erlaubt die Berücksichtigung zweier verschiedener Variationsebenen : die der Verschiedenheit innerhalb der Stämme (W) und die der Unterschiede zwischen den Stämmen (B). Die Kenntnis der Basen W und B erlaubt die Bestimmung weiterer, besonders interessanter Variablen :

sie enthalten jegliche zur Unterscheidung der Ausgangsvariablen nötigen Informationen,

sie liefern die besten Unterscheidungsmöglichkeiten,, und es besteht keine Korrelation zwischen ihnen.

So kann die Information ohne Verlust verdichtet werden, und man braucht sich nur an die für die Unterscheidung wichtigen Faktoren zu halten. Die determinierten Faktoren — es sind dies lineare Kombinationen von Ausgangsvariablen — legen die Verwendung neuer, besonders unterscheidbarer Variablen, wie gewisser Varianz-Koeffizienten, nahe.

Diese ersten Ergebnisse müssen durch eine Klassifizierung der Stämme ergänzt werden, d. h. gewisse, sich genügend nahestehende Stämme werden in ein und derselben Gruppe zusammengefasst, um eine neue Population zu bilden. Auf diese Weise werden die Stämme nach und nach in einer Art Baum — einem Dendogramm — vereinigt, wodurch ein synthetisches Bild der phaenotypischen Beziehungen zwischen den Stämmen gewonnen wird. Die Methoden der numerischen Taxonomie, die dazu dienen, die Klasseneinteilung festzulegen, sind durch sehr allgemeine Vorstellungen von der Gruppenbildung, wie Menge, Dichte und Netz charakterisiert. Diese Vorstellungen werden beschrieben, zwei davon angewandt. Mit ihrer Hilfe lassen sich die einzelnen Stämme auf fast die gleiche Weise klassifizieren, und man erhält so die Bestätigung der gewonnenen Klasseneinteilung.

Die durch das Dendogramm festgelegte Einteilung in Abbildung 2 wird in Tabelle 2 erläutert. Der Ökotyp « Provence », der ganz auf der linken Seite des Dendogramms angesiedelt ist, bildet eine Gruppe, die sich wesentlich von den anderen Ökotypen der Honigbiene *Apis mellifica mellifica* unterscheidet, und zwar mit einer Maximaldistanz von 4,77. Ferner findet man die Ökotypen « Bretagne » und « Essonne » (Pariser Region) mit $D^2 = 6{,}82$ und den Ökotyp « Landes » mit $D^2 = 10{,}59$.

Die verschiedenen Gruppen von Kreuzungen zwischen *Apis mellifica mellifica* und *Apis mellifica ligustica* sind dem Dendrogramm mit $D^2 = 12{,}19$ und $D^2 = 13{,}40$ angefügt.

Die reinrassigen *Apis mellifica ligustica* haben $D^2 = 35{,}12$ und 38,18; die reinrassigen *Apis mellifica intermissa* $D^2 = 58{,}34$.

Auch die Einordnung neuer Stämme kann in betracht gezogen werden : ihre Klassifizierung erfordert die Einführung eines Masses für die Distanz zwischen den Stämmen. Diese Distanz kann für einen neuen Stamm durch Vergleich mit allen bereits analysierten Distanzen ermittelt werden, und man erhält auf diese Weise eine objektive Einordnung.

Alle Berechnungen sind mit in Fortran IV geschriebenen Programmen durchgeführt worden, die zur Verfügung stehen.

TABLEAU 2. — *Détail des 50 souches étudiées, dans l'ordre établi par le dendrogramme*

TABELLE 2. — *Einzelangaben der 50 untersuchten Stämme in der im Dendrogramm festgelegten Reihenfolge*

N° des souches pour l'ordinateur Bezeichnung d. Stämme f. d. Datenverarbeitung	N° d'origine des souches Bezeichnung d. Herkunft d. Stämme	Ecotype ou Race Ökotyp oder Rasse	Année de création Jahrgang	GROUPES PRINCIPAUX HAUPTGRUPPEN
6	86	P	1961	
8	280	P	1961	
7	89	P	1961	1961
10	S - 305	P	1961	
9	338	P	1961	
12	S - 903	P	1962	
15	t - 15	P	1962	
24	t - 24	P	1964	
13	S - 904	P	1962	1962
17	F - S - 901	P	1963	1963
14	S - 901	P	1962	1964
16	P - 17	P	1963	
18	F - S - 902	P	1963	
20	F - S - 905	P	1964	
21	F - S - 905	P	1964	
11	F - S - 902	P	1962	
19	t - 19	P	1963	
3	261	P	1960	
5	S - 901	P	1960	
4	327	P	1960	1960
1	17	P	1960	
2	S - 902	P	1960	
29	S - 902	P	1966	
30	F - S - 905 x S - 901	P	1966	
34	t - 34	P	1966	
39	t - 39	P	1967	
906	S - 905 x S - 901	P	1967	
31	F - S - 901 x S - 904	P	1966	1965
903	S - 903	P	1967	1966
25	F - S - 904	P	1965	1967
902	S - 902	P	1967	
907	S - 904	P	1967	
38	F - S - 905 x S - 902	P	1967	
901	S - 901	P	1967	
900	S - 900	P	1967	
26	F - S - 902	P	1965	
28	t - 28	P	1965	
905	S - 905	P	1967	
701	0	B	1966	
700	0	E	1967	écotypes B.E.L.
702	0	L	1966	
23	A. lig. (Elbe) x S - 905	H	1964	
32	A. lig. x S - 902	H	1966	hybrides
36	A. lig. x S - 901	H	1967	A. mel. mellifica
35	A. lig. x S - 905	H	1967	A. mel. ligustica
37	A. lig. x S - 902	H	1967	
33	0	I	1966	
802	0	I	1967	A. mel. ligustica
800	0	I	1964	
801	0	M	1963	A. mel. intermissa

écotype Provence — Apis mellifica mellifica

RÉFÉRENCES BIBLIOGRAPHIQUES

CACOULOS T., 1965. Comparing Mahalanobis distances I : comparing distances between known normal populations and another unknown. *Sankhya, série A*, **27** (1), 1-22.

FRESNAYE J., 1965. Étude biométrique de quelques caractères morphologiques de l'abeille noire française *(Apis mellifica mellifica)*. *Ann. Abeille*, **8** (4), 271-283.

GIAVARINI I., 1953. Ricerche sui caratteri razziali, dell'*Apis mellifica ligustica spinola*. *Estr. Soc. Ent. Ital.*, **32**, 119-128.

GOETZE G., 1940. *Die beste Biene*. 200 p. Liedloff, Loth et Macharlis Leipzig.

GOETZE G., 1963. *Die Honigbiene in natürlicher und künstlicher Zuchtauslese*. 212 p. Paul Parey, Hambourg.

LOUIS J., LEFEBVRE J., MORATILLE E., FRESNAYE J., 1968. Essai de discrimination de lignées consanguines d'abeilles domestiques *(Apis mellifica mellifica* L.*)* obtenues par insémination artificielle *C. R. Acad. Sci., Paris*, t. **267**, 526-528.

MILLIER C., TOMASSONE R., 1969. Méthodes d'ordination et de classification : leur efficacité et leurs limites. *Coll. Inter. C.N.R.S.* : emploi des calculateurs en archéologie, Marseille, 207-228.

MYINT TIN, 1965. Comparison of some ratio estimators., *J.A.S.A.*, *60*, 309, 294-307.

RAO C. R., 1965. *Linear Statistical inference and its applications*. John Wiley, New York, 522 p.

ROUX M., 1968. *Un algorithme pour construire une hiérarchie particulière*. Thèse 3e cycle, Paris.

RUTTNER F., 1952. Die Aussenmerkmale der Carnica-Stammes Troiseck *Öster. Imker*, *2* (4). 67-69.

RUTTNER F. et MACKENSEN O., 1964. The genetics of the honeybee. *Bee World*, **33**, 53-62 71-79.

RUTTNER F., 1963. *Die Zuchtauslese bei der Biene*, N.O. Imkerschule Wr. Neustadt, Walthergasse 6, 79 p.

RUTTNER F. 1968. *Les races d'abeilles. Traité de Biologie de l'abeille, tome I*, Masson édit. Paris, 27-44.

SNEATH P. H. A., 1968. Evaluation of clustering methods. *Coll. Num. Taxon*, St Andrews, 216-225.

SOKAL R. R. et SNEATH P. H. A., 1963. *Principles of numerical taxonomy*. Freeman, San Francisco, 359 p.

Vanden DRIESSCHE R., 1965. La recherche des constellations de groupes à partir des distances généralisées. D^2 de Mahalanobis, *Biom Praxim*, 6, 36-47.

WISHART D., 1968. Mode analysis : a generalisation of nearest neighbour which reduces chaining effects. *Coll. Num. Taxon*, St Andrews, 233-254.

86	n° de souche sans descendance = Nr. des Stammes ohne Nachkommen
S - 905	n° de souche conservée = Bezeichnung d. weitergeführten Stammes
t - 15	groupe de colonies témoins = Testgruppen
F - S - 902	souche fille d'une souche conservée = Tochterstamm eines weitergeführten Stammes
P	*Apis mellifica mellifica* écotype Provence = Apis mellifica mellifica Ökotyp Provence
B	*Apis mellifica mellifica* écotype Bretagne = Apis mellifica mellifica Ökotyp Bretagne
E	*Apis mellifica mellifica* écotype Essonne = Apis mellifica mellifica Ökotyp Essonne
L	*Apis mellifica mellifica* écotype Landes = Apis mellifica mellifica Ökotyp Landes
I	*Apis mellifica ligustica* (Italie) = Apis mellifica ligustica (Italien)
M	*Apis mellifica intermissa* (Maroc) = Apis mellifica intermissa (Marokko)
H	colonies d'abeilles hybrides = Volker mit Hybridbienen