

L'EXPLORATION DES DONNÉES GÉOGRAPHIQUES SUR REGARD

Anthony UNWIN*

RÉSUMÉ Le logiciel REGARD offre des outils graphiques interactifs d'exploration des données spatiales. Son utilisation est illustrée ici par un ensemble de données sur l'environnement et la reproduction des oiseaux en Écosse.

ABSTRACT REGARD is a software package which provides high interaction graphics tools for exploring spatial data. Its use is illustrated using a data set on bird breeding and environment in Scotland.

RESUMEN El programa REGARD proporciona instrumentos gráficos interactivos de exploración de los datos espaciales. Se ilustra aquí su utilización con un conjunto de datos sobre el medio ambiente y la reproducción de las aves en Escocia.

• DONNÉES SPATIALES • ÉCOSSE • REGARD • REPRODUCTION DES OISEAUX

• BIRD BREEDING • REGARD • SCOTLAND • SPATIAL DATA

• DATOS ESPACIALES • ESCOCIA • REGARD • REPRODUCCIÓN DE LAS AVES

Il est difficile d'analyser les données géographiques à partir de modèles statistiques traditionnels pour des raisons tenant aux interdépendances spatiales, à la complexité des structures spatiales, au nombre élevé des variables et au fait que beaucoup de données relèvent de l'observation de terrain. Les géographes ont à leur disposition tout un éventail de cartes pour présenter leurs données de manière efficace (Brunet, 1987), mais ces cartes statiques sont rarement capables de représenter plus de deux variables à la fois. Les outils graphiques interactifs développés à des fins statistiques peuvent être adaptés à l'exploration des données géographiques à condition que le lien soit établi entre les données statistiques et leur localisation géographique. C'est ce qui a été fait avec le logiciel *REGARD* (*Radical Effective Graphical Analysis of Regional Data — Analyse graphique effective et radicale des données régionales*), pour micro-ordinateur Macintosh.

La principale caractéristique de *REGARD* est l'établissement de liens entre les plans d'affichage de cartes ponctuelles, aréales ou linéaires, et les représentations statistiques des données associées aux objets cartographiés, ce qui permet l'exploration interactive des caractéristiques des zones sélectionnées (de la carte vers les données) ou de localisations géographiques présentant des propriétés statistiques particulières (des données vers la carte). Le fonctionnement se fait naturellement dans les deux sens: un ensemble de zones peut se détacher parce qu'elles présentent

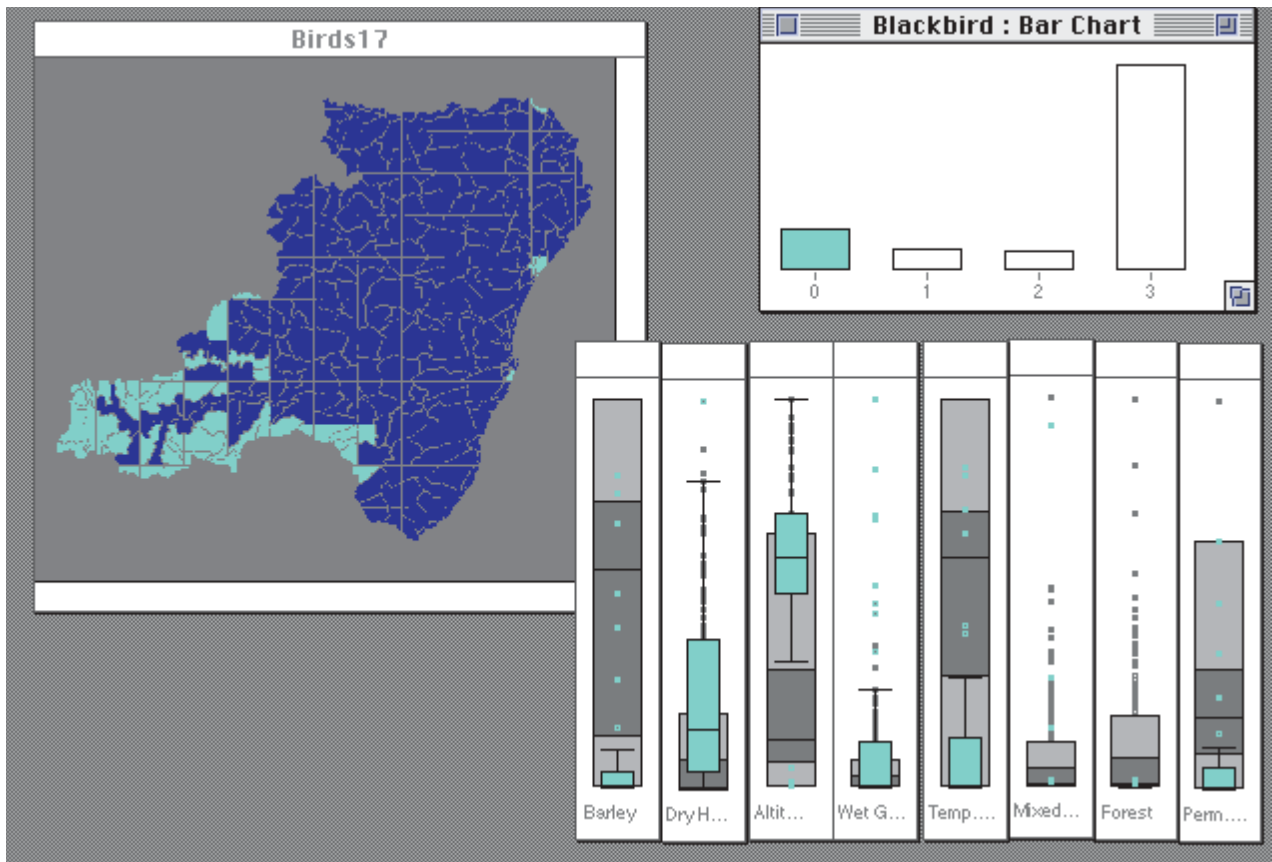
toutes une valeur élevée (par exemple toutes les zones connues de reproduction du lagopède rouge d'Écosse, *red grouse*), mais on peut aussi explorer un autre groupe de variables pour rechercher les caractéristiques environnementales éventuelles distinguant cet ensemble de zones du reste de la région.

REGARD a été conçu comme un outil d'aide à l'analyse des données spatiales. Son but est de permettre aux experts d'un domaine d'interagir directement avec leurs données. On trouvera des précisions et des exemples supplémentaires dans plusieurs articles (Unwin *et al.*, 1990; Wills *et al.*, 1991 et Unwin *et al.*, 1992). Parmi les autres développements, qui sont rares dans ce secteur, on peut signaler le travail de Nagel, utilisant le logiciel *ISP* (Nagel *et al.*, 1991).

Les oiseaux du Nord-Est de l'Écosse

De 1981 à 1984, des ornithologues écossais ont collecté un grand nombre de données sur la reproduction de plus de 100 espèces d'oiseaux. Ils ont divisé la région en près de 400 sites d'environ 10 km² chacun, dont les contours correspondaient à l'habitat naturel des oiseaux (Buckland *et al.*, 1990). Un certain nombre de variables décrivant l'environnement de chaque site ont été retenues. Le présent article présente un sous-ensemble de ces données, comprenant 15 variables environnementales et des données sur la reproduction de 35 espèces d'oiseaux. Les catégories de reproduction ont été définies respectivement comme certaine (3), probable (2), possible (1), données inexistantes (0).

* Institut de Mathématiques, Université d'Augsburg, Allemagne.



1. Affichage d'une variable: zones où le merle ne semble pas se reproduire

Traduction de la légende

Birds 17, Oiseaux 17 — Blackbird: Bar Chart, Merle: diagramme en bâtons

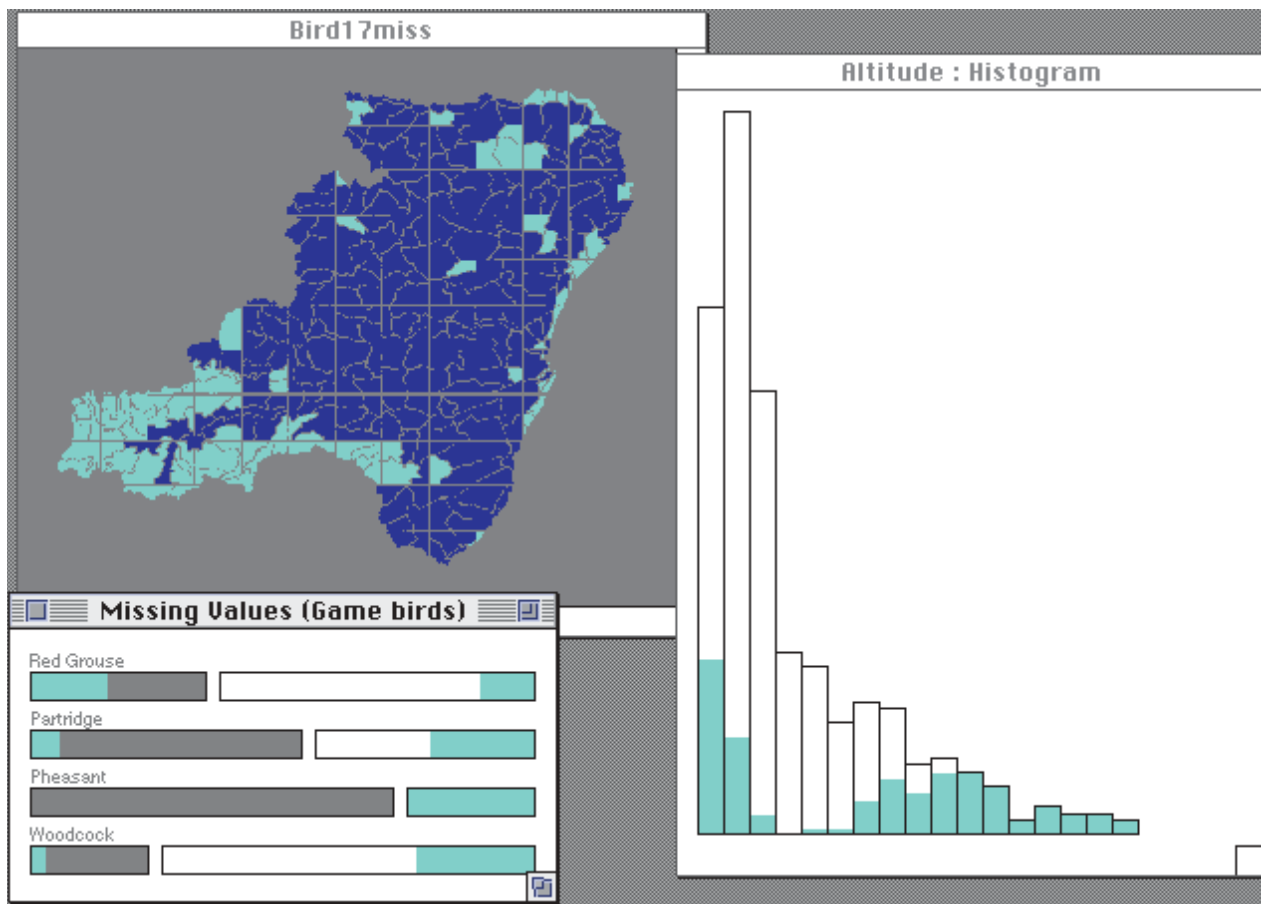
Barley, Orge — Dry H..., Lande sèche — Altit..., Altitude — Wet G..., Prairie humide — Temp..., Température — Mixed..., Forêt mixte — Forest, Forêt — Perm..., Prairie permanente.

Exemples d'utilisation de *REGARD*

La figure 1 montre les zones où le merle (*blackbird*) ne semble pas se reproduire. Sur le diagramme en bâtons représentant sa reproduction, on a choisi de sélectionner le bâton 0. Il apparaît qu'il ne se reproduit ni dans la région-frontière du Sud-Ouest ni dans certaines régions littorales. La figure 1 comporte aussi des diagrammes en boîtes représentant une sélection de variables environnementales. Le diagramme classique en «boîte et moustaches» représente les données pour les zones sélectionnées, et la boîte d'arrière-plan les données concernant toutes les zones. Ces dernières peuvent avoir pour caractéristiques des altitudes supérieures à celles des autres zones, des températures plus basses, moins de cultures d'orge, moins de forêts, plus de lande sèche et moins de prairies permanentes. On notera aussi que certaines variables ont une distribution dissymétrique, ce qui semble indiquer que des transformations sont nécessaires. *REGARD* peut afficher ensemble de nombreuses variables grâce à des boîtes reliées entre elles, ce qui occupe moins de place que des histogrammes et beaucoup moins encore que des graphiques bivariés. Il faut interpréter avec prudence la présence d'un grand nombre

de graphiques interdépendants comme ceux qui représentent des variables environnementales. Il est nécessaire, chaque fois que cela est possible, de confirmer les résultats d'analyses graphiques par des méthodes d'analyse plus traditionnelles.

La figure 2 présente le graphique d'un groupe de régions comparable à celui de la figure 1, mais obtenu par des méthodes tout à fait différentes. Certaines régions ont été repérées comme n'étant pas de celles où se reproduit une certaine espèce d'oiseaux: le code adopté pour cette espèce particulière est alors cette fois non pas 0 mais •, c'est-à-dire le code correspondant aux données manquantes. Cela a permis d'utiliser la représentation spécifique disponible dans *REGARD* pour les valeurs manquantes (fig. 2). Pour chaque variable sélectionnée (ici quatre espèces de gibier à plumes — *game birds*), la barre de gauche, colorée, montre les zones pour lesquelles des valeurs existent, tandis que la barre de droite, blanche, fait apparaître les zones à valeurs inexistantes. On a sélectionné les zones où aucun signe de reproduction du faisan (*pheasant*) n'a été enregistré. L'intérêt de cet affichage est de permettre la comparaison de l'extension spatiale de plusieurs variables.



2. Affichage de plusieurs variables: zones de reproduction de quatre espèces

Traduction de la légende

Birds 17miss..., Oiseaux 17 valeurs manquantes — *Missing Values (Game birds)*, Valeurs manquantes (gibier à plumes) — *Red Grouse*, Lagopède rouge — *Partridge*, Perdrix — *Pheasant*, Faisan — *Woodcock*, Bécasse des bois — *Altitude: histogram*, Altitude: histogramme.

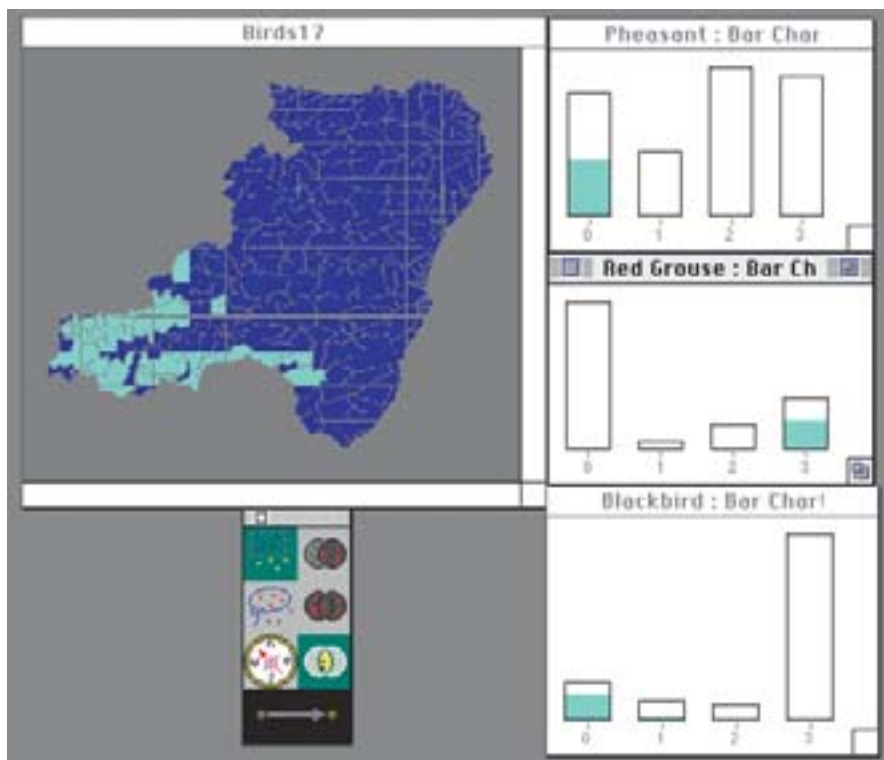
La synthèse ainsi obtenue des informations sur la reproduction des quatre espèces est de toute évidence bien plus économique que les quatre tableaux de légendes qui auraient été nécessaires dans un système différent. Il est manifeste que le lagopède tend à se reproduire dans des zones différentes de celles du faisan; il peut donc être intéressant, dans une étape ultérieure, de comparer les zones où les deux espèces se reproduisent avec celles où un seul est présent. Il est clair d'après la carte que les zones de non-reproduction du faisan comprennent à la fois le littoral et le Sud-Ouest (qui est une région de collines, comme le montre l'analyse précédente): cela ne fait pas un groupe homogène. C'est pourquoi on n'a pas utilisé de diagramme en boîte; l'histogramme de la figure 2 confirme d'ailleurs la validité de ce choix, puisque deux groupes ressortent clairement, à savoir les zones les plus élevées et quelques-unes des plus basses. Cet exemple prouve l'intérêt qu'il y a à visualiser les données statistiques de plusieurs façons différentes, chaque type de représentation apportant des informations différentes.

La figure 3 reprend l'analyse du faisan et du lagopède pour illustrer l'utilisation de l'outil de sélection inter-affichages. La palette

d'outils en bas à gauche utilise comme icônes des diagrammes de Venn pour ses options de sélection. Dans le cas présent, la plus basse a été privilégiée, indiquant que c'est le mode d'intersection qui est utilisé. Au départ, c'était le mode «par défaut» (l'option tout en haut) qui avait été utilisé, et les zones sélectionnées étaient celles où le faisan ne se reproduit pas (le bâton 0). Les caractéristiques mises en valeur étaient par conséquent les mêmes que dans la figure 2. Puis on a choisi le mode «intersection» pour sélectionner le bâton indiquant la catégorie de reproduction «certaine» (3) dans la série concernant le lagopède. Seules ressortent les zones où les deux conditions sont réunies: elles se concentrent dans le Sud-Ouest. On a utilisé le même diagramme en bâtons pour la reproduction du merle dans la figure 1, ce qui permet de constater que ces mêmes zones sont celles où il est peu probable que le merle se reproduise.

Quels autres outils ajouter?

REGARD a été conçu d'un point de vue statistique plutôt que géographique. L'éventail des fonctionnalités géographiques disponibles est par conséquent limité. Il serait utile d'y ajouter des



3. Illustration de l'utilisation de l'outil de sélection inter-affichages

Traduction de la légende

Birds 17, Oiseaux 17 — *Pheasant: Bar Char...*, Faisan: diagramme en bâtons — *Red Grouse: Bar Ch...*, Lagopède rouge: diagramme en bâtons — *Blackbird: Bar Chart*, Merle: diagramme en bâtons.

outils permettant de définir des nombres de voisins (définis de diverses manières), de calculer des superficies et des distances, d'agréger des données, d'explorer des limites. On pourrait incorporer dans *REGARD* des fonctions définies par l'utilisateur, et certains outils de ce genre ont déjà été testés.

Du fait de leur rapidité et de leur souplesse, les outils graphiques interactifs sont des instruments très puissants. Il est plus difficile en revanche de garder en mémoire les voies d'exploration et les analyses préliminaires; bien que ce problème se pose dans de nombreux logiciels, il est toutefois moins gênant dans le cas de programmes gouvernés par un langage de commande puisque, s'il ne parvient à aucun résultat, l'utilisateur peut toujours refaire tout son itinéraire depuis le début. Il faudrait pouvoir disposer d'un moyen simple d'afficher les cartes correspondant aux diverses étapes d'une analyse afin d'établir des comparaisons: par exemple pour comparer plus facilement les cartes des figures 1, 2 et 3. On peut utiliser à cette fin l'un des nombreux utilitaires de capture d'écran disponibles, mais on ne peut conserver en mémoire que les cartes elles-mêmes, et non le cheminement par lequel elles sont obtenues.

Data Desk (Velleman, 1992) et *REGARD* proposent tous deux des outils de sélection graphique permettant à l'utilisateur de

faire des sélections multivariées par des méthodes graphiques plutôt que par des formules SQL ou autres procédures analytiques. C'est là une manière d'interroger les cartes qui semble naturelle, et l'on apprécierait de pouvoir disposer de facilités encore accrues.

Par ailleurs, il est souvent intéressant de comparer des groupes de zones par des mesures sommaires effectuées sur l'ensemble des groupes (par exemple les zones où les oiseaux se reproduisent ou non). Pour cela, un pas en avant très important serait la création d'outils rapides et souples permettant de former ces groupes, et d'afficher des illustrations efficaces facilitant les comparaisons. Aucun autre logiciel n'offre cette possibilité. En effet, si beaucoup de logiciels statistiques permettent de produire des illustrations et des tableaux comparatifs, ces derniers sont très lourds et les illustrations rarement interactives: il s'agit davantage d'attirer l'attention sur le logiciel que sur le processus d'analyse lui-même. C'est pourquoi *REGARD* permet sans doute de

voir les cartes d'un œil nouveau...

Références bibliographiques

- BRUNET R., 1987, *La Carte mode d'emploi*, Paris, Fayard/Reclus, 270 p.
- BUCKLAND S. T., BELL M. V. et PICOZZI N., 1990, *The Birds of North-East Scotland*, Aberdeen, North-East Scotland Bird Club.
- NAGEL M., HUBER T. M. et HOERING H., 1992, «Hochinteraktive Graphik und ihre Anwendung auf Daten mit territorialem Bezug», in: Faulbaun F. (éd.), *Sofstat 91*, Stuttgart, G. Fischer, pp. 177-184.
- UNWIN A. R., WILLS G. et HASLETT J., 1990, «*REGARD* - Graphical Analysis of Regional Data», *ASA Proceedings of the Section on Statistical Graphics*, pp. 36-41.
- UNWIN A. R., SLOAN B. J. et WILLS G., 1992, «Interactive Graphical Methods for Trade Flows», *Proceedings of Conference on New Techniques and Technologies*, Bonn, pp. 295-303.
- VELLEMAN P. F., 1992, *Data Desk, Data Description*, Ithaca NY.
- WILLS G., UNWIN A. R. et HASLETT J., 1991, «Spatial Interactive Graphics Applied to Irish Socio-economic Data», *Proceedings of 1991 ASA Statistical Graphics Section*, pp. 37-41.

L'auteur remercie S. T. Buckland, actuellement professeur de statistique à l'université d'Aberdeen, qui a bien voulu lui fournir ses données.