Évaluer les prévisions probabilistes de population

Evaluating Probabilistic Population Forecasts

Nico Keilman*

Résumé – Les statisticiens ont développé des règles de notation pour évaluer les prévisions probabilistes par rapport aux observations. Toutefois, on en trouve peu d'applications dans la littérature sur les prévisions de population. Une règle de notation mesure la distance entre la distribution prédictive et le résultat. Nous passons en revue les règles de notation qui privilégient l'exactitude (le résultat est proche de l'espérance de la distribution) et la précision (la distribution présente une faible variance, de sorte qu'il est difficile d'atteindre l'objectif). Nous évaluons les prévisions de population probabilistes établies pour la France, les Pays-Bas et la Norvège. Les prévisions de la taille de la population totale des Pays-Bas et de la Norvège ont obtenu de bons scores. L'erreur sur la population de base a engendré un mauvais score pour la prévision française. Nous évaluons aussi la prévision de la composition par âge et par sexe pour 2010. Les prédictions relatives aux Pays-Bas ont reçu les meilleurs scores, excepté celles concernant les personnes très âgées. Pour la Norvège, le score de la structure par âge reflète la sous-prédiction de l'immigration après l'élargissement de l'Union européenne en 2005.

Abstract – *Statisticians have developed scoring rules for evaluating probabilistic forecasts against observations. However, there are very few applications in the literature on population forecasting. A scoring rule measures the distance between the predictive distribution and its outcome. We review scoring rules that reward accuracy (the outcome is close to the expectation of the distribution) and sharpness (the distribution has low variance, which makes it difficult to hit the target). We evaluate probabilistic population forecasts for France, the Netherlands, and Norway. Forecasts for total population size for the Netherlands and for Norway performed quite well. The error in the jump-off population caused a bad score for the French forecast. We evaluate the age and sex composition predicted for the year 2010. The predictions for the Netherlands received the best scores, except for the oldest old. The age pattern for the Norwegian score reflects the under-prediction of immigration after the enlargement of the European Union in 2005.*

Codes JEL / JEL classification : C15, C44, J11
Mots-clés : prévisions de population probabilistes, règle de notation, modèle par cohorte et composante
Keywords: probabilistic population forecast, scoring rule, cohort component model

* *Département d'économie de l'Université d'Oslo, Norvège (nico.keilman@econ.uio.no)*
Remerciements – L'auteur exprime sa reconnaissance à Laurent Toulemon et à trois rapporteurs anonymes pour leurs excellents commentaires.

Reçu en mars 2019, accepté en février 2020.
Traduit de la version originale anglaise

Citation: Keilman, N. (2020). Evaluating Probabilistic Population Forecasts. *Economie et Statistique / Economics and Statistics*, 520-521, 49–64. https://doi.org/10.24187/ecostat.2020.520d.2033

La plupart des instituts de statistique qui effectuent des prévisions de population recourent à une approche déterministe (NRC, 2000). Ils analysent les tendances historiques en matière de fécondité, de mortalité et de migration, puis extrapolent ces tendances pour le futur, en mobilisant des avis d'experts et des techniques statistiques. Ces extrapolations reflètent leurs meilleures anticipations. En plus de calculer l'évolution probable de la taille et de la structure de la population, de nombreux instituts calculent également une variante haute et une variante basse de sa future croissance, afin d'attirer l'attention des utilisateurs sur l'incertitude qui entoure les prévisions démographiques. Par exemple, les précédentes projections de population officielles en France tablaient sur 76.5 millions d'habitants en 2070 si les tendances à l'œuvre se poursuivent (Blanpain & Buisson, 2016). Toutefois, la croissance effective jusqu'en 2070 pourrait être plus faible ou plus forte que les tendances actuelles ne le suggèrent, débouchant sur une population comprise entre 66.1 millions et 87.6 millions de personnes selon des hypothèses de trajectoires hautes et basses pour la fécondité (1.8 ou 2.1 enfants par femme en moyenne après 2020), pour l'espérance de vie des hommes (entre 87.1 et 93.1 ans en 2070) et celle des femmes (entre 90 et 96 ans) et pour la migration internationale (excédent de migration compris entre 20 000 et 120 000 personnes par an).

L'un des inconvénients majeurs de cette approche déterministe est de ne pas quantifier l'incertitude. La probabilité de recenser entre 66.1 et 87.6 millions d'habitants en France en 2070 est-elle de 30 %, de 60 % ou de 90 % ? Pourtant, à des fins de planification, les utilisateurs ont fréquemment besoin de connaître le degré de confiance qu'ils peuvent accorder aux chiffres prédits. Quelle devrait être la robustesse du système de retraite face à une augmentation rapide ou lente de l'espérance de vie ? Devons-nous prévoir des capacités supplémentaires dans les écoles primaires, au cas où les naissances seraient beaucoup plus nombreuses que prévu ? Comme Keyfitz (1981) l'écrivait il y a près de quarante ans : « Les démographes ne peuvent pas plus être tenus responsables de l'inexactitude des prévisions de population sur un horizon de vingt ans que les géologues, les météorologues et les économistes qui ne savent pas prédire les tremblements de terre, la rigueur des hivers ou les dépressions vingt ans à l'avance. Ce que nous devons faire, c'est nous avertir les uns les autres, ainsi que le public, quant à la probabilité d'erreur inhérente à nos estimations ». Pour cette raison, certains instituts de statistique ont commencé à publier leurs prévisions sous forme de distributions de probabilités, suivant des pratiques courantes dans les domaines de la météorologie et de l'économie, entre autres. Le Centraal Bureau voor de Statistiek (CBS, l'institut de statistique des Pays-Bas) est pionnier en la matière (voir Alders & De Beer, 1998). Statistics New Zealand (2011) en Nouvelle-Zélande et l'Istituto Nazionale di Statistica (ISTAT, 2018) en Italie les publient également. À cet égard, il convient de citer la Division de la population des Nations Unies, qui est chargée de la mise à jour à intervalles réguliers des prévisions de population de l'ensemble des pays. En 2014, la Division a publié la première série officielle de prévisions de population probabilistes pour tous les pays, à l'aide de la méthodologie développée par Raftery *et al.* (2012)[1]. L'objectif d'une prévision probabiliste n'est pas d'estimer des tendances futures qui soient plus exactes que les prévisions déterministes, mais de fournir à l'utilisateur une image plus complète de l'incertitude des prévisions.

Les instituts de statistique pourraient mettre en œuvre des nouvelles méthodes et des travaux développés par les démographes et les statisticiens depuis les années 1980. Deux développements méritent d'être mentionnés. Le premier, l'approche par simulation. L'approche analytique s'appuie sur un modèle stochastique par cohorte et composante, dans lequel les distributions statistiques relatives à la fécondité, à la mortalité et à la migration sont transformées en distributions statistiques relatives à la taille de la population et à sa structure par âge et par sexe, ce qui nécessite des hypothèses solides, sans quoi les expressions du moment de second ordre de la distribution par âge et par sexe restent approximatives. Aujourd'hui, l'approche par simulation, répandue, évite les hypothèses simplificatrices et les approximations propres à l'approche analytique. L'idée est de calculer plusieurs centaines ou milliers de variantes de prévision (« parcours d'échantillonnage ») en fonction d'une sélection aléatoire des valeurs des paramètres d'entrée pour la fécondité, la mortalité et la migration. Les résultats des simulations sont stockés dans une base de données. Keilman (2009) donne un exemple pour la France. Deuxième changement méthodologique récent : le passage d'une approche principalement fréquentiste à une vision bayésienne de la probabilité. Dans l'approche fréquentiste, la probabilité d'un événement est

1. Voir également http://esa.un.org/unpd/wpp/Graphs/Probabilistic/POP/TOT/

liée à la fréquence relative à laquelle il survient. Dans l'approche bayésienne en revanche, une probabilité est interprétée comme étant l'opinion subjective du statisticien, ce qui est particulièrement utile lorsque les modèles reposent sur des opinions d'experts et lorsque l'on combine ce type d'informations avec des données. Le passage d'une approche fréquentiste à une approche bayésienne en matière de prévisions de population s'inscrit dans une tendance plus générale vers la « démographie bayésienne », qui a commencé à prendre de l'essor il y a une dizaine d'années (Bijak & Bryant, 2016). Les prévisions probabilistes des Nations Unies, que nous avons citées, sont un bon exemple de l'approche bayésienne. Costemalle (ce numéro) applique la méthode pour la France.

L'exactitude d'une prévision probabiliste ne peut être évaluée que dix à vingt ans après sa publication, une fois les données observées *ex post facto* concernant la taille et la structure par âge de la population disponibles. Mais l'exercice reste difficile car il nécessite de comparer les probabilités prédites par le prévisionniste avec les probabilités réelles mais inconnues des événements étudiés. Pour cette raison, les statisticiens ont développé des « règles de notation », également appelées « fonctions de notation ». Une règle de notation mesure la distance entre la distribution prédite d'une variable démographique et sa valeur réelle (Gneiting & Raftery, 2007 ; Gneiting & Katzfuss, 2014). Le score obtenu pour une variable donnée n'a pas de signification intrinsèque. L'interprétation des scores n'est utile que dans le cadre d'une comparaison, ce qui explique pourquoi les fonctions de notation sont fréquemment utilisées pour comparer deux prévisions probabilistes alternatives.

Bien que la méthodologie de l'évaluation des prévisions probabilistes et des règles de notation soit connue depuis un moment déjà, elles ont été peu appliquées aux projections de population. Shang *et al.* (2016) ont évalué l'exactitude des prévisions probabilistes par cohorte et par composantes au Royaume-Uni et comparé les deux méthodes, en utilisant une règle de notation pour les intervalles de prédiction. Shang (2015) et Shang & Hyndman (2017) ont évalué les prévisions par intervalles pour les taux de mortalité par âge dans divers pays, puis utilisé des scores d'intervalle pour choisir les meilleures méthodes de prévision de la mortalité. Alexopoulos *et al.* (2018) ont appliqué des scores d'intervalle aux intervalles de prédiction des taux de mortalité par âge en Angleterre, au Pays de Galles et en Nouvelle-Zélande, puis évalué la performance prédictive de cinq modèles de prédiction de la mortalité. Ces quatre articles évaluent les prévisions démographiques probabilistes à partir d'échantillons partiels : les paramètres sont estimés à partir des années les plus anciennes, et les prédictions du modèle sont confrontées aux données disponibles les plus récentes. À notre connaissance, une évaluation *ex post* de prévisions de population probabilistes n'a jamais été tentée par le passé.

Cet article vise à montrer comment des méthodes d'évaluation des prévisions probabilistes développées dans d'autres domaines peuvent être appliquées aux prévisions de population. Nous présentons et appliquons des règles de notation aux intervalles de prédiction, ainsi qu'à des échantillons simulés de la taille et de la structure par âge de la population projetée. À l'aide de données relatives à la France, aux Pays-Bas et à la Norvège, nous présentons les règles de notation puis nous comparons les prévisions probabilistes calculées par différents chercheurs, avec trois objectifs. Le premier est d'analyser la rapidité avec laquelle l'exactitude d'une prévision probabiliste change en fonction du délai de réalisation, c'est-à-dire lorsqu'elle se rapproche de l'horizon de projection. Le deuxième est de comparer la précision de deux prévisions probabilistes (alternatives) pour un même pays. Le troisième est d'analyser la performance relative des prévisions entre différents pays.

La section 1 examine comment les résultats d'une prévision probabiliste sont mis à disposition : en tant qu'intervalles de prédiction ou au moyen d'une base de données. La section 2 présente plusieurs règles de notation et leurs caractéristiques. La section 3 donne des exemples empiriques. Avant de conclure, nous évaluons diverses prévisions probabilistes de la taille de la population totale et de la pyramide des âges de la population de trois pays.

1. Publier une prévision de population probabiliste

Les méthodes utilisées pour évaluer une prévision probabiliste dépendent fortement de la façon dont les résultats de cette prévision sont mis à disposition. Il y a principalement deux possibilités : l'une consiste à publier des intervalles de prédiction pour les variables de la population, l'autre consiste à fournir aux utilisateurs une base de données contenant les parcours d'échantillonnage.

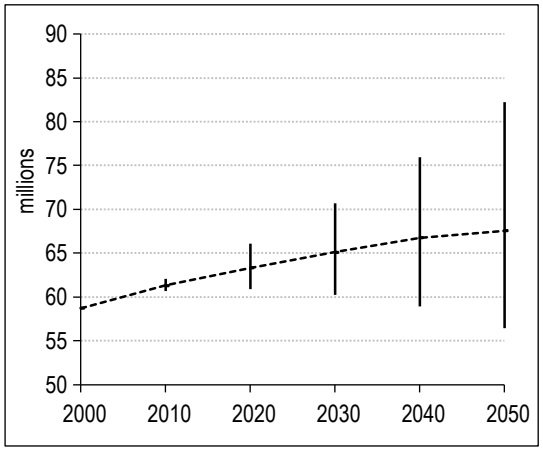
Costemalle (ce numéro) présente des intervalles de prédiction de la population pour la France, calculés selon une approche bayésienne. Par exemple, une probabilité de 80 % pour que

cette population se situe entre 68.1 millions et 75.0 millions de personnes en 2070 (voir la figure XV de son article). L'auteur présente également des intervalles de prédiction de 95 %, qui couvrent des situations plus extrêmes. D'autres chercheurs (voir les exemples de la section 3) donnent des prévisions probabilistes avec des intervalles de prédiction de 67 %.

La figure I montre des intervalles de prédiction de 80 % pour la population de la France, tirés du projet « Uncertain Population of Europe » (UPE). L'année de base de cette prévision probabiliste est l'année 2003. En 2050 (47 ans plus tard), l'intervalle de prédiction de 80 % est de 25.7 millions de personnes (82.2 – 56.5), beaucoup plus large que celui de Costemalle, de 6.9 millions de personnes (75.0 – 68.1, après 46 ans). Différentes perceptions de l'incertitude des prédictions relatives aux futurs taux de fécondité, de mortalité et de migration internationale engendrent des intervalles de prédiction plus précis (optimisme) ou plus larges (pessimisme).

Les auteurs utilisent donc des intervalles de prédiction ayant différentes probabilités de couverture. Une probabilité de 67 % ou de 80 % couvre la plupart des prévisions mais exclut la queue de distribution des erreurs, plus volatile. Les auteurs qui utilisent une probabilité de couverture de 95 % sont certainement influencés par les sciences sociales où construire des intervalles de confiance à 95 % ou procéder à une vérification des hypothèses avec une probabilité faible (5 %) pour les erreurs de type I (c'est-à-dire en rejetant une hypothèse nulle quand bien même elle serait vraie) est courant. En revanche, un intervalle de prédiction assorti d'une probabilité de couverture de 67 % ou de 80 % donne à l'utilisateur une idée de l'écart potentiel à la valeur ponctuelle prédite, ce qui est très différent de la construction d'intervalles de confiance et de la vérification d'hypothèses. Nous utiliserons des intervalles de prédiction de 67 % et de 80 % dans la section 3.

Figure I – **Valeurs médianes et intervalles de prédiction de 80 % pour la population totale de la France métropolitaine**



Note : les valeurs médianes sont représentées par le trait en pointillés et les intervalles de prédiction par les traits pleins.
Lecture : la probabilité que la population soit inférieure à 67.7 millions en 2050 se chiffre à 50 % et celle que la population soit supérieure à 67.7 millions est la même. La probabilité que la population totale atteigne entre 56.5 millions et 82.2 millions en 2050 se chiffre à 80 %.
Source : Keilman (2009).

Les intervalles de prédiction ne sont qu'un résumé de la distribution de probabilité complète de la variable concernée. Dans certains cas on peut supposer que la distribution sous-jacente est approximativement normale. Il est alors possible de déduire ses paramètres à partir des bornes supérieure et inférieure de l'intervalle. Toutefois, certaines variables de population se limitent à une partie de la forme réelle, comme la part des personnes âgées dans la population (entre 0 et 1). L'hypothèse de normalité de la distribution n'est alors pas adéquate, et nous perdons un gros volume d'information en ne publiant que des intervalles de prédiction, et pas les distributions sous-jacentes.

La plupart des informations ne sont mises à disposition qu'une fois que toutes les trajectoires simulées sont stockées dans une base de données, que l'utilisateur peut consulter (Alho & Spencer, 2005). Un exemple courant est le jeu de prévisions de population probabilistes établies pour 18 pays européens dans le cadre du projet UPE. Le modèle par cohorte et composantes a été appliqué 3 000 fois pour chaque pays, avec une population de base déterministe (au 1er janvier 2003) et des valeurs spécifiques à l'âge variant de façon probabiliste pour la fécondité, la mortalité et la migration nette. L'horizon des prévisions est l'année 2050. Les prévisions du projet UPE sont intéressantes pour deux raisons. La première, l'un des buts explicites est de quantifier l'incertitude de façon à ce qu'elle reflète la volatilité historique de la fécondité, de la mortalité et de la migration internationale. La deuxième, le projet est le premier à avoir examiné de façon exhaustive la corrélation empirique des erreurs de prévision en matière de fécondité, de mortalité et de migration dans différents pays. Le site Web du projet[2] contient une base de données rassemblant les résultats de simulations ($N = 3\,000$) pour les hommes et les femmes par tranches d'âge de cinq années, pour des intervalles de temps de dix ans (2010(10)2050), et ce pour chaque pays. L'utilisateur peut établir son ou ses propres

2. *http://www.stat.fi/tup/euupe/index_en.html. On y trouve aussi plusieurs articles publiés et non publiés.*

histogrammes pour une ou plusieurs variables d'intérêt. Dans la section 3, nous utiliserons les prévisions de pyramides des âges établies pour 2010 en France, aux Pays-Bas et en Norvège afin d'illustrer les règles de notation examinées à la section 2.

2. Évaluation

Soit X la variable pour laquelle nous calculons une prévision, dont la fonction de distribution cumulative (FDC) est définie comme $F(x) = \mathrm{P}(X \leq x)$. La fonction de densité de probabilité (FDP) de X est $f(x) = \frac{dF(x)}{dx}$. Nous supposons toujours l'existence des intégrales et des différents moments de la distribution de probabilité. Pour une analyse détaillée s'appuyant sur la théorie des probabilités, voir par exemple Gneiting & Katzfuss (2014) et Gneiting & Raftery (2007). Soit y la valeur observée de X. Une fonction de notation $S(F(x),y)$ attribue une valeur numérique (« score ») à la prévision $F(x)$, compte tenu de l'observation y. $S(F(x),y)$ prend les valeurs de la droite réelle $\mathbb{R}$ (incluant éventuellement plus et moins l'infini).

Le postulat suivant est un bon point de départ pour définir une fonction de notation : une prévision qui prédit le résultat réel avec une forte probabilité doit recevoir un bon score. Il « fonctionne » bien pour les prévisions catégorielles, lorsque X est une variable aléatoire discrète. Toutefois, s'agissant de prévisions de nombres de personnes (par âge, par sexe et par année de prévision), X s'apparente davantage à une variable aléatoire continue qu'à une variable aléatoire discrète (sauf si la prévision porte sur une population de très petite taille). Dans la suite de l'article, nous supposons que la prévision et la fonction de notation s'appliquent à une variable aléatoire continue. De nombreuses fonctions de notation sont construites à partir des deux principes suivants. Premièrement, une observation proche de la médiane ou de l'espérance de la distribution prédictive engendre un bon score – plus elle est proche et mieux c'est. La règle de notation est alors sensible à la distance (Staël von Holstein, 1970 ; Murphy, 1970). Deuxièmement, compte tenu d'une observation donnée, une distribution prédictive étroite (« précise ») engendre un bon score – plus elle est étroite et mieux c'est. Par exemple, un intervalle de prédiction de 80 % couvrant une observation donnée constitue une meilleure prévision qu'un intervalle de 67 % aussi large couvrant la même observation, car il est relativement difficile d'atteindre l'objectif lorsque la variance de la FDP est faible. Toutefois, les deux principes n'ont pas la même importance. On peut arguer que, si l'observation est « trop loin » de la médiane ou de l'espérance, une FDP étroite ne devrait plus être bien notée. En d'autres termes, si le prévisionniste « prend un risque » (c'est-à-dire prédit une FDP étroite), la prévision devrait obtenir un bon score lorsqu'elle est proche de la médiane ou de l'espérance mais ne devrait pas obtenir un bon score lorsqu'elle en est trop loin. La signification de « trop loin » n'est pas claire et varie d'une règle de notation à l'autre. Dans l'exemple ci-dessus, cela signifie que « l'observation sort de l'intervalle de prédiction ». Ce choix peut être critiqué car il repose sur une nette dichotomie. Dans un très petit intervalle aux alentours de la borne supérieure ou de la borne inférieure de l'intervalle de prédiction, la prévision passe brusquement d'un bon score à une pénalité si elle se situe juste en dehors de l'intervalle. En d'autres termes, compte tenu de la distribution prédictive et de la valeur observée, un intervalle de prédiction dont la borne inférieure est légèrement inférieure à la valeur observée engendre un bon score, tandis qu'un intervalle de prédiction dont la borne inférieure est légèrement supérieure à la valeur observée engendre un mauvais score. Les probabilités de couverture sont arbitraires (on utilise souvent 80 % mais 81 % ou 79 % fonctionnent également très bien). Pour cette raison, nous devons être prudents avec la notion de « trop loin ».

Certaines des règles de notation que nous examinons ci-après reposent sur l'idée que la proximité est plus importante que la précision. Toutefois, comme nous le verrons, le sens que nous donnons à « trop loin » diffère selon les règles de notation. D'autres règles de notation considèrent que les deux principes sont indépendants. On dira qu'une fonction de notation est orientée négativement lorsqu'un score inférieur implique une meilleure prévision, et qu'elle est orientée positivement dans le cas contraire. En conséquence, une fonction de notation orientée négativement peut être interprétée comme une pénalisation, et une fonction de notation orientée positivement comme une récompense.

De nombreuses règles de notation différentes ont été suggérées, en fonction de la nature de la prévision. On trouvera dans Gneiting & Raftery (2007) et Jordan *et al.* (2019) une vue d'ensemble exhaustive sur la question. Nous nous limiterons aux règles de notation applicables aux variables aléatoires continues. Une catégorie de règles de notation s'applique aux prévisions de densité en fonction d'expressions de forme close de la FDC ou de la FDP. Citons par exemple le score logarithmique $\mathrm{LogS}(F(x), y) = -\log(f(y))$. Une autre

catégorie de règles de notation, plus adaptée au sujet de cet article, évalue des échantillons simulés. Dans ce cas, la distribution prédictive n'est pas disponible sous forme analytique. Une deuxième différence est celle qui existe entre les prévisions d'une variable et les prévisions de plusieurs variables. Pour ces dernières, tant la variable prédite X que l'observation y se composent d'un vecteur. Jordan *et al.* (2019) ont développé le package R 'scoringRules' qui couvre de nombreuses situations dans le cadre de travaux appliqués.

Nous présentons ci-dessous trois types de règles de notation : celles basées uniquement sur les deux premiers moments de la distribution prédictive (section 2.1), celles découlant de la simulation de la distribution prédictive complète, fournie en tant qu'échantillon (section 2.2) et enfin celles pour lesquelles nous ne disposons que des intervalles de prédiction (section 2.3).

2.1. Fonctions de notation basées sur la variance

Supposons une FDP unimodale de la prévision. Lorsque le résultat réel est proche du centre de la densité prédite (caractérisée par la moyenne, la médiane ou le mode), cette prévision est meilleure que celle dont le résultat est loin du centre. En d'autres termes, la prévision obtient un meilleur score lorsque X présente peu de variation autour de y que lorsque la variation est plus marquée. Cela conduit à une fonction de notation basée sur la variance, que nous nommons « *VS* » (*Variance-based Scoring*) dans le reste de cet article et que nous définissons comme suit.

Soit VS la variance de X autour de la valeur observée y, où

$$VS = \int (x - y)^2 \, f(x) \, dx \qquad (1)$$

Pour y égal à l'espérance de X (que nous écrivons μ), VS réduit la variance de X, que nous écrivons σ^2. L'expression (1) donne

$$VS = \sigma^2 + (\mu - y)^2 \qquad (2)$$

Cela définit une fonction de notation simple basée sur la variance, qui pourrait servir à évaluer la qualité de la FDP prédictive unimodale. Gneiting & Raftery (2007) la citent comme une fonction de notation qui correspond au critère de choix du modèle prédictif (*predictive model choice criterion*, PMCC). Nous pouvons l'appliquer pour les fonctions de densité analytiques et pour les échantillons simulés. Pour ces derniers, on utilise les valeurs de σ^2 et de μ estimées à partir de l'échantillon. Cette fonction de notation est orientée négativement : un score inférieur indique une meilleure prévision. Elle récompense à la fois l'exactitude (lorsque y coïncide avec μ, la prévision est de qualité optimale) et la précision (une faible variance engendre un bon score, que la prévision soit proche ou non).

Pour une prévision déterministe (ponctuelle), σ^2 est égal à zéro et la prévision est μ. Dans ce cas, VS diminue jusqu'au niveau de l'erreur quadratique de la prévision. Les erreurs de ce type sont à la base de l'erreur quadratique moyenne fréquemment utilisée dans l'évaluation des prévisions de population déterministes (Alho & Spencer, 2005 ; Smith *et al.*, 2001 et Keilman, 1990).

Une autre fonction de notation, le score de Dawid-Sebastiani (*DSS*), se fonde également sur les deux premiers moments de la distribution prédictive (voir par exemple Gneiting & Katzfuss, 2014).

$$DSS = \ln(\sigma^2) + (\mu - y)^2/\sigma^2 \qquad (3)$$

Cette fonction de notation est semblable au score VS basé sur la variance de l'expression (2), mais donne un poids différent à la variance de la prévision σ^2.

Une faible variance engendre un bon score (bas) tant que $\frac{dDSS}{d\sigma^2} = \frac{1}{\sigma^2} - \frac{(\mu - y)^2}{\sigma^4} > 0$, ou $\sigma > |\mu - y|$. Tandis que le VS récompense toujours les distributions prédictives présentant une faible variance, le DSS ne le fait que si l'observation y s'éloigne de l'espérance de la distribution prédictive de moins d'un écart type.

Imaginons un prévisionniste qui sait que sa prévision probabiliste sera, en temps voulu, évaluée au moyen de la règle de notation (2) ou (3). Supposons qu'à un certain stade du processus de production de la prévision, le problème soit d'étalonner le modèle de prévision. Selon la règle de notation (2) ou (3), cet étalonnage devrait s'attacher en priorité à choisir une valeur appropriée pour l'espérance μ de la distribution prédictive – non pas de la médiane ou de tout autre paramètre de position. De fait, il existe une corrélation étroite entre l'étalonnage du modèle et l'évaluation de la prévision. Si la situation est claire lorsqu'il n'y a qu'un seul utilisateur, elle est plus complexe lorsque les utilisateurs sont nombreux et lorsque leurs règles de notation sont différentes (ou inconnues).

2.2. Le score de probabilité *CRPS*

Le *CRPS* (*continuous ranked probability score*) peut servir de score standardisé pour évaluer la prévision probabiliste de variables à valeur

réelle (Gneiting & Raftery, 2007). Il est défini en fonction de la FDC prédictive $F(x)$ comme

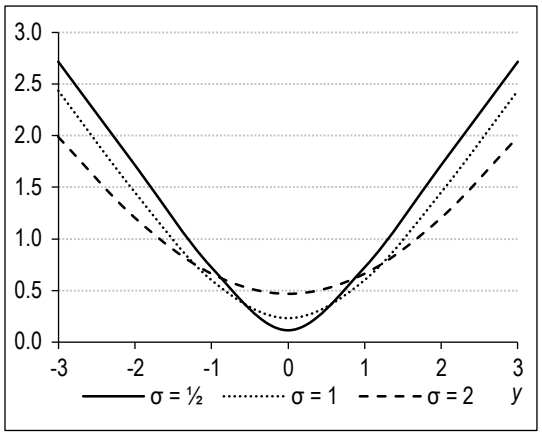
$$CRPS(F,y) = \int \left(F(z) - \mathbb{I}\{y \le z\}\right)^2 dz \qquad (4)$$

où $\mathbb{I}\{y \le z\}$ désigne la fonction caractéristique, égale à 1 si $y \le z$ et à zéro dans le cas contraire. La forme précise du *CRPS* vient du score de Brier (1950). Le score de Brier, ou score de probabilité, est l'erreur quadratique moyenne d'une prévision de probabilité catégorielle. Murphy (1970) l'a adapté aux catégories ordonnées pour X, donnant lieu au *RPS* (*ranked probability score*). Matheson & Winkler (1976) ont proposé un *RPS* pour les variables aléatoires continues, le *CRPS*.

Les solutions triviales à l'équation (4) sont rares. Jordan *et al.* (2019) dressent la liste des cas recensés. Par exemple, lorsque $F(z)$ est la distribution normale standardisée $\Phi(.)$ avec une densité $\varphi(.)$, le $CRPS(\Phi,y)$ est égal à $y(2\Phi(y) - 1) + 2\varphi(y) - 1/\sqrt{\pi}$. La distribution normale, avec une espérance μ et un écart type σ, donne $\sigma CRPS(\Phi,(y-\mu)/\sigma)$.

Quelques exemples concrets permettent d'illustrer le *CRPS*. Prenons une distribution normale et supposons, sans perte de généralité, que μ est égal à zéro. La figure II trace le *CRPS* en fonction de y, c'est-à-dire qu'elle représente sa sensibilité à la distance. Nous présentons trois cas, à savoir des écarts type de ½, de 1 et de 2. Par construction, μ étant égal à 0, les courbes sont symétriques aux alentours de zéro. Comme attendu, le meilleur score est obtenu lorsque y est égal à zéro. Le score se dégrade lorsque la valeur absolue de y augmente, c'est-à-dire lorsque y est loin de μ. La précision de la FDP prédictive (écart type faible) n'est récompensée qu'au sein d'un certain intervalle y aux alentours de zéro. Par exemple, une prévision parfaite (y égal à zéro) obtient un meilleur score pour σ = ½ ($CRPS$ = 0.1168) que pour σ = 2 ($CRPS$ = 0.4674). Toutefois, la FDP contenant σ = 2 obtient un meilleur score que celle contenant σ = ½ pour les observations y dont la valeur absolue est supérieure à environ 0.9. L'intervalle dans lequel la précision est récompensée est plus court pour les valeurs σ faibles que pour les valeurs élevées.

Figure II – ***CRPS* pour une distribution normale, avec une espérance μ égale à zéro et des observations y allant de -3 à +3**



Source : calculs de l'auteur.

Les prévisions de population probabilistes sont habituellement calculées en tant que distributions simulées et l'on ne peut pas calculer l'intégrale de l'expression (4). Dans ce cas, il est utile de partir du principe selon lequel (4) peut s'écrire ainsi :

$$CRPS(F,y) = E_F|X_1 - y| - \tfrac{1}{2}E_{F,F}|X_1 - X_2| \qquad (5)$$

où X_1 et X_2 sont des variables aléatoires indépendantes avec une distribution F (Gneiting & Raftery, 2007). Le *CRPS* mesure la proximité de l'observation y à laquelle on peut s'attendre par rapport à la variable prédite X, corrigée de la distance attendue entre toutes les paires de valeurs possibles de X. Cette distance prédite est faible lorsque l'écart type de F est faible. Toutes choses égales par ailleurs, une augmentation de l'écart type engendre un meilleur score. Toutefois, lorsque l'écart type change, la première espérance $E_F|X_1 - y|$ change également. Cette règle de notation récompense-t-elle toujours la précision, ou seulement dans un certain intervalle ? Cela reste une question empirique.

Le *CRPS* réduit l'erreur absolue lorsque F est une prévision déterministe. Supposons une prévision disponible en termes de distribution simulée. Dans le cas, la FDC est $\widehat{F}_m(x) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\{X_i \le x\}$ où m est la taille de l'échantillon, et (5) devient

$$CRPS\left(\widehat{F}_m, y\right) = \frac{1}{m}\sum_{i=1}^{m}|X_i - y| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|X_i - X_j|.$$

La mise en œuvre de cette expression est inefficiente car son ordre computationnel est $o(m^2)$. Une représentation plus efficiente, et algébriquement équivalente, est (Jordan *et al.,* 2019, p. 6)

$$CRPS\left(\widehat{F}_m, y\right) = \sum_{i=1}^{m}(X_{(i)} - y)\left(m\mathbb{I}\{y < X_{(i)}\} - i + \tfrac{1}{2}\right) \qquad (6)$$

où $X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(m)}$, est l'échantillon simulé ordonné. Le *CRPS* est défini dans l'expression (6) comme étant toujours positif, car chaque terme de la somme est positif.

2.3. Scores d'intervalle

De nombreuses prévisions de population probabilistes sont présentées en tant que prévisions par intervalles et non pas en tant que distributions de probabilité (simulées) – voir section 1. Considérons un intervalle de prédiction central $(1–\alpha)$, avec des extrémités inférieure et supérieure correspondant aux quantiles prédictifs aux niveaux $\alpha/2$ et $(1–\alpha/2)$ respectivement[3]. Écrivons l et u pour les quantiles inférieur et supérieur. Gneiting & Raftery (2007) définissent la fonction de notation suivante :

$$(u-l)+\frac{2}{\alpha}\left[(l-y)\mathbb{I}\{y<l\}+(y-u)\mathbb{I}\{y>u\}\right] \quad (7)$$

En prenant α, le score d'intervalle de Gneiting-Raftery (*Gneiting-Raftery interval score* – ci-après *GRIS*) récompense les prévisions pour des intervalles de prédiction étroits qui saisissent l'observation y : lorsque deux prévisions contradictoires présentent des intervalles de prédiction différents pour un α donné, la prévision dont l'intervalle de prédiction est le plus court obtient le meilleur score (le plus bas). En revanche, si la valeur de y est située en dehors de l'intervalle de prédiction, nous obtenons un mauvais score (plus élevé). La pénalité appliquée à la sortie de l'intervalle de prédiction est plus importante pour un α faible que pour un α élevé. Le *GRIS* peut facilement être appliqué à l'intervalle de prédiction d'une variable avec des délais de réalisation différents : un an à l'avance, deux ans à l'avance, trois ans à l'avance, etc.

Le *GRIS* ne récompense pas toujours la précision, même lorsque l'intervalle saisit correctement la réalisation. Supposons deux prévisions alternatives ayant le même intervalle de prédiction $[l,u]$ mais des probabilités de couverture différentes. Par exemple, une prévision donne une probabilité de 67 % à l'intervalle de prédiction $[l,u]$ tandis que l'autre donne une probabilité de couverture de 80 % à ce même intervalle. La deuxième prévision est plus précise et devrait recevoir un meilleur score lorsque l'observation y reste dans les limites de $[l,u]$. Mais cela n'est pas le cas, parce que le *GRIS* est indépendant de α dans cette situation. Pour régler ce problème, nous pouvons utiliser une version légèrement modifiée du *GRIS*, à savoir

$$GRISmod=\alpha(u-l)+\beta\left[(l-y)\mathbb{I}\{y<l\}+(y-u)\mathbb{I}\{y>u\}\right] \quad (8)$$

où $\beta > 0$ est un paramètre qui détermine la rapidité avec laquelle le score se détériore lorsque l'observation s'éloigne soit de la borne supérieure soit de la borne inférieure de l'intervalle de prédiction. Une valeur β élevée engendre une pénalité plus importante qu'une valeur faible. Le *GRISmod* récompense la précision tant pour une valeur α fixe et des intervalles de prédiction différents que pour un intervalle de prédiction fixe et des valeurs α différentes. Lorsque β est égal à deux, le *GRISmod* est égal à α*GRIS.* Si la valeur β est égale à la probabilité α*,* le *GRISmod* diminue jusqu'au niveau de $\alpha(u–y)$ pour $y < l$ et $\alpha(y–l)$ lorsque $y > u$.

Au lieu d'utiliser des fonctions de notation pour les intervalles de prédiction, nous pourrions vérifier la fréquence à laquelle les données réelles tombent dans les limites des intervalles. Par exemple, Raftery *et al.* (2012) ont validé leur méthode bayésienne de prévision de la population de 159 pays en estimant le modèle rassemblant les données d'une période de quarante ans (1950-1990), afin de générer une distribution prédictive pour la totalité de la population par âge et par sexe pour une période de vingt ans (1990-2010). Ils ont ensuite comparé les distributions des intervalles de prédictions de 80 % et de 95 % qui en résultent avec les observations réelles, puis ont vérifié la proportion de l'échantillon de vérification tombant dans les limites de leurs intervalles. Ces proportions étant proches des valeurs nominales de 80 % et de 95 %, les auteurs ont conclu à la validité de leur approche. Cette méthode présente un gros inconvénient : elle compare les données et les intervalles de prédiction de nombreuses variables, comme la taille de la population des 56 pays d'Afrique à un moment donné. Toutefois, les corrélations régionales de la fécondité, la mortalité et/ou la migration suggèrent que les tailles des populations des 56 pays ne sont pas indépendantes. On dispose de moins de données qu'escompté initialement et les proportions observées ne peuvent pas être comparées directement aux valeurs nominales (Alho & Spencer 2005, p. 248).

2.4. Fonctions de notation utilisées dans les applications empiriques

Dans la section 3, nous utilisons le *CRPS* de l'expression (6) pour évaluer des prévisions pour lesquelles nous disposons de résultats de simulation détaillés. Si nous n'avons que des intervalles de prédiction, nous utilisons le score *VS* basé sur la variance de l'expression (2), le

3. Nous supposons que les deux quantiles sont connus. Si l'on veut évaluer les prévisions par intervalles lorsque le niveau de couverture nominal est précisé mais que les quantiles sur lesquels les intervalles sont fondés ne sont pas précisés, l'approche présentée ici ne peut pas être appliquée (Askanazi et al.*,2018).*

score de Dawid-Sebastiani (*DSS*) de l'expression (3) et les scores d'intervalle (*GRIS* et *GRISmod*) des expressions (7) et (8). Pour le *GRISmod*, nous supposons que la valeur du paramètre β est égale à la probabilité α utilisée pour définir l'intervalle. Le *VS* et le *DSS* utilisent l'espérance et l'écart type de la distribution prédictive. Dans la mesure où seules les bornes supérieure et inférieure de l'intervalle sont disponibles, nous supposons la normalité et définissons l'espérance comme étant la moyenne des deux bornes. Parallèlement, nous estimons l'écart type comme la moitié de la largeur de l'intervalle pour les intervalles de 67 % et la largeur de l'intervalle divisée par 2.564 pour les intervalles de 80 %.

Notons que le score dépend de l'échelle de la variable X pour laquelle nous disposons d'une distribution prédictive (qui correspond à l'échelle de l'observation y). Par conséquent, lorsque nous comparons les scores de deux prévisions pour des pays dont les populations sont de tailles très différentes, la population la plus petite reçoit le meilleur score, quel que soit son exactitude. Pour que la comparaison soit juste, nous devons tenir compte de la taille de la population. Nous avons normalisé le *VS*, le *DSS*, le *CRPS*, le *GRIS* et le *GRISmod* comme suit :

- le *VS* est divisé par μ^2, c'est-à-dire le carré de l'espérance de la distribution prédictive ;

- le *DSS* est normalisé en soustrayant $2ln(\mu)$ de sa valeur initiale[4] ;

- le *CRPS*, le *GRIS* et le *GRISmod* sont divisés par μ.

3. Résultats

Nous illustrons les règles de notation présentées à la section 2.4 en évaluant les prévisions de population probabilistes de trois pays : la France, les Pays-Bas et la Norvège. Nous nous concentrons sur la taille de la population totale (section 3.1) et sur la pyramide des âges (section 3.2). Les données proviennent de sources variées :

1. Le site Web du projet UPE (voir section 1) fournit des échantillons ($N = 3\,000$) pour les prévisions de pyramides des âges des trois pays pour les années 2010, 2020, …, 2050. Nous utilisons les résultats de 2010.

2. Alho & Nikander (2004) présentent des intervalles de prédiction de 80 % et des médianes de la taille de la population totale, entre autres, pour chaque année de la période 2004-2050 pour tous les pays du projet UPE. Nous utilisons les résultats de 2004-2019.

3. Pour les Pays-Bas, nous avons des informations sur les prévisions de population probabilistes officielles, en prenant l'année 2000 comme année de base (voir CBS, 2001). Les données contiennent des intervalles de prédiction de 67 % et des espérances de population totale, pour chaque année de la période 2000-2050, ainsi que pour les hommes et les femmes répartis entre différentes tranches d'âge de cinq années, pour des intervalles de cinq ans.

4. Pour la Norvège, nous utilisons les résultats du projet de projections stochastiques « StocProj » (Keilman *et al.,* 2002) dont le but était de calculer une prévision de population probabiliste en prenant l'année 1996 comme année de base. Comme nous ne disposons pas des résultats détaillés de ces simulations, nous utilisons à la place des intervalles de prédiction de 80 % pour la taille de la population totale pour les années 1997-2019.

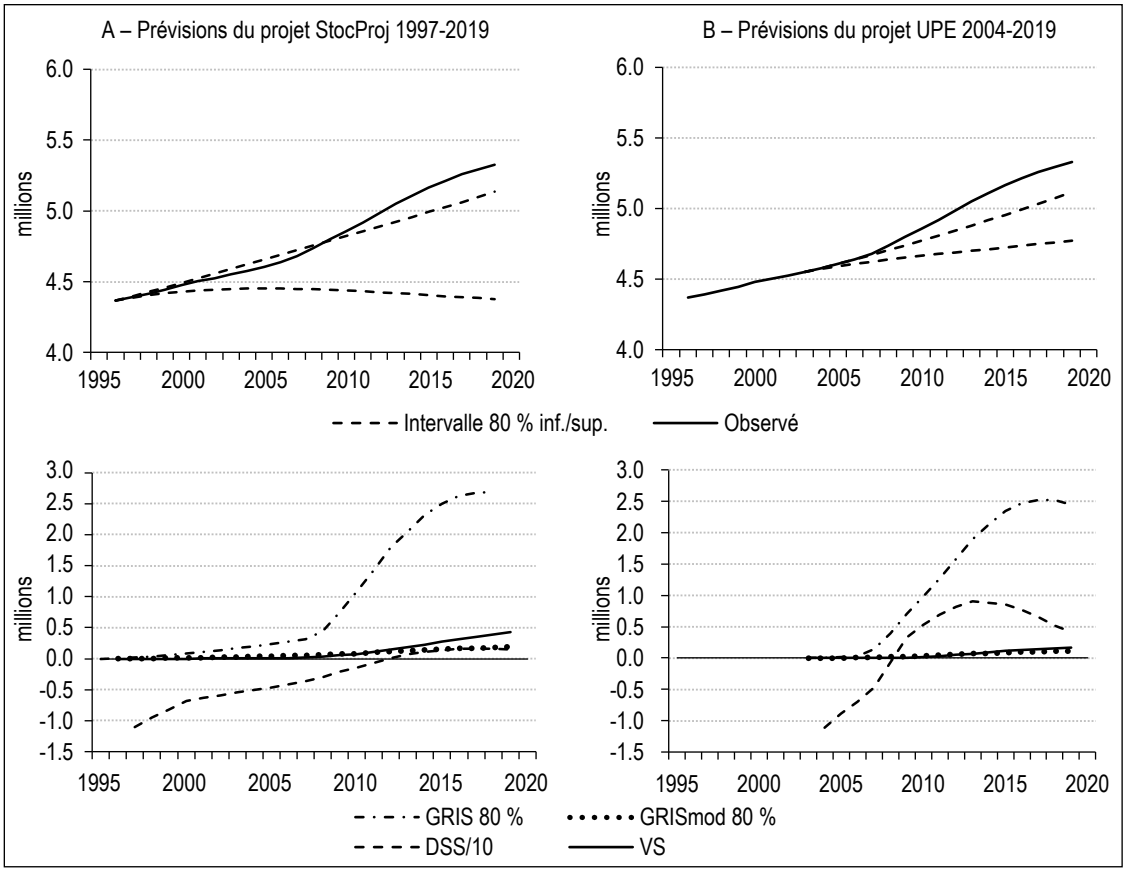
3.1. Taille de la population

La figure III illustre nos résultats pour la Norvège. Nous présentons quatre graphiques : deux pour le projet StocProj (gauche) et deux pour les prévisions du projet UPE (droite). Les deux graphiques du haut donnent des intervalles de prédiction de 80 % et des valeurs observées pour la taille de la population totale, tandis que les deux graphiques du bas présentent les scores des deux prévisions.

Les deux prévisions sous-estiment la population totale à partir d'environ 2005. Cela vient principalement du fait que, après l'élargissement de l'Union européenne, les travailleurs migrant des pays baltes et d'Europe de l'Est vers la Norvège ont été beaucoup plus nombreux que prévu. À noter que, pour chaque délai de réalisation prévu, les intervalles de prédiction du projet StocProj sont plus larges que ceux du projet UPE. Le score d'intervalle modifié *GRISmod* récompense la précision et est donc inférieur – et par conséquent meilleur – pour le projet UPE que pour le projet StocProj, même si la différence est minime (cf. lignes en pointillé). Le score d'intervalle modifié *GRISmod* et le score basé sur la variance *VS* affichent la même tendance : les deux prévisions se détériorent progressivement à mesure que le délai de réalisation augmente. Les courbes en tirets correspondent au score de Dawid-Sebastiani *DSS* divisé par dix, afin

4. L'intérêt réside dans la valeur du DSS *pour une variable aléatoire à l'échelle* X/N *avec une valeur à l'échelle* y/N *de* y *(*N *non aléatoire et positif), que nous écrivons* DSS(y/N)*. Alors* DSS(y/N) = $2ln(\sigma/N) + [(\mu/N - y/N)/(\sigma/N)]^2$ = DSS(y) – 2ln(N)*. Pour* N*, nous avons choisi l'espérance de taille de la population* μ.

Figure III – **Taille de la population totale de la Norvège**
Intervalles de prédiction, valeurs observées et scores

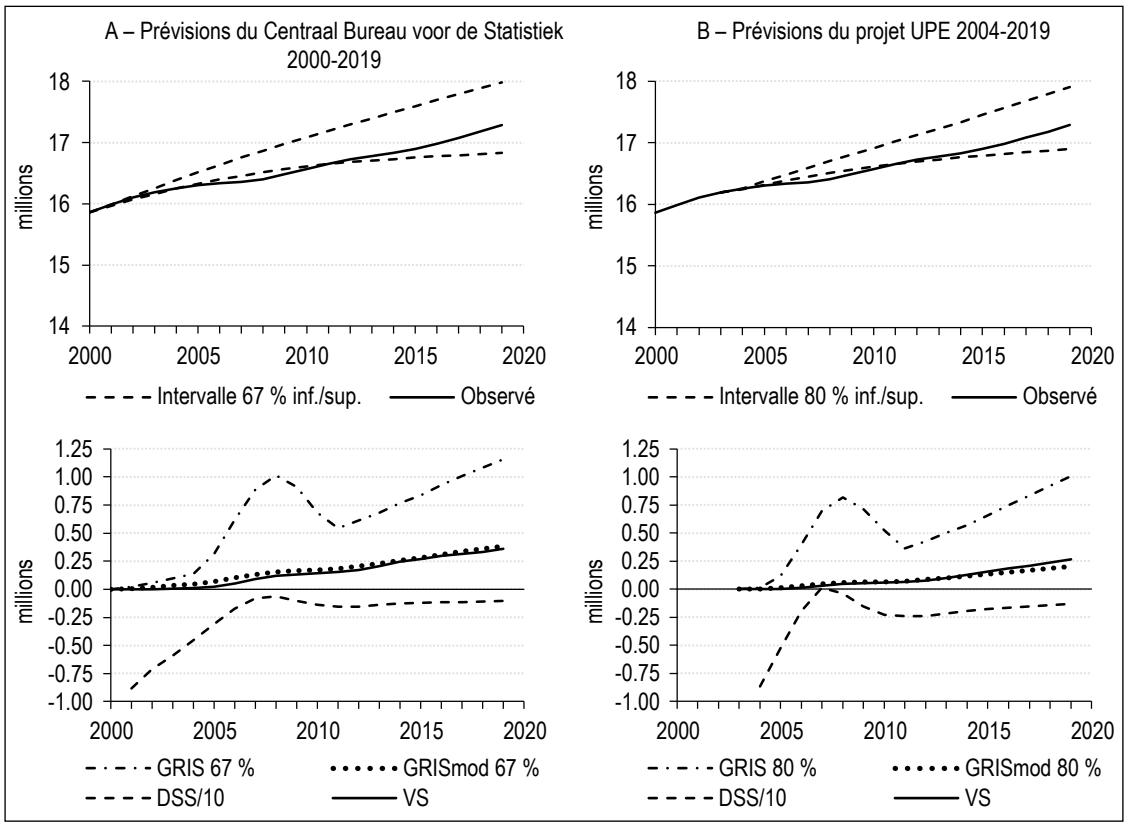


Note : les intervalles de prédiction et les valeurs observées sont présentés dans la partie supérieure, tandis que les scores d'intervalle (*GRIS* et *GRISmod*), le score de Dawid-Sebastiani (*DSS*) et le score basé sur la variance (*VS*) sont présentés dans la partie inférieure. Les intervalles de prédiction, les valeurs observées et les scores *GRIS, GRISmod* et *VS* sont exprimés en millions. Le score de Dawid-Sebastiani est divisé par dix.
Source : voir premiers paragraphes de la section 3.

qu'il puisse être tracé sur le même graphique que les trois autres scores. Le *DSS* commence par des valeurs négatives dans les deux cas, car l'écart type σ des deux prévisions de taille de la population est faible (mesuré en millions) durant les premières années. Par exemple, $\sigma = 0.0039$ dans le projet StocProj pour 1997, ce qui donne $\ln(\sigma^2) = -11.1162$. Puisque $((\mu\text{-}y)/\sigma)^2 = 0.0309$, le *DSS* est égal à – 11.0853 et tracé à hauteur de – 1.1085 dans la figure III. Le *DSS* augmente fortement dans le projet UPE car il ne récompense plus la précision dès lors que la valeur observée s'écarte de plus d'un écart type de l'espérance ($|\mu - y| > \sigma$, voir section 3.1). C'est le cas pour chaque année pour laquelle nous avons les données du projet UPE, c'est-à-dire à partir de 2004. Pour le projet StocProj, le cas d'un écart type trop faible pour récompenser la précision ne survient pas avant 2008, soit douze ans dans le futur. En revanche, les fonctions de notation *GRISmod* et *VS* ne pénalisent pas les prévisions « trop optimistes » (c'est-à-dire celles pour lesquelles la variance de la distribution prédictive est trop petite). À noter que, pour le projet StocProj, le *DSS* se stabilise à partir d'environ 2016, soit vingt ans dans le futur.

Pour les prédictions de la taille de la population totale des Pays-Bas, les intervalles de prédiction de 80 % du projet UPE reflètent une prévision plus précise que les intervalles de 67 % de la prévision de CBS (figure IV). Dans les deux cas, la taille observée de la population sort des intervalles pendant plusieurs années jusqu'en 2011. Ensuite, les observations reviennent dans les limites des intervalles. Le score d'intervalle modifié de la prévision du projet UPE est bien meilleur que celui de la prévision de CBS. Les scores d'intervalle ignorent le fait que les valeurs observées se rapprochent du centre des intervalles, dans la mesure où ces scores excluent les informations relatives à la moyenne, à la médiane ou au mode de la distribution prédictive. À l'aune des scores de Dawid-Sebastiani, les deux prévisions sont de même qualité. Dans les deux cas, le *DSS* se stabilise à partir de 2010. Cela vient du fait que l'erreur de prévision $|\mu - y|$

Figure IV – **Taille de la population totale des Pays-Bas. Intervalles de prédiction, valeurs observées et scores**



Note : les intervalles de prédiction et les valeurs observées sont présentés dans la partie supérieure, tandis que les scores d'intervalle (*GRIS* et *GRISmod*), le score de Dawid-Sebastiani (*DSS*) et le score basé sur la variance (*VS*) sont présentés dans la partie inférieure. Les intervalles de prédiction, les valeurs observées et les scores *GRIS, GRISmod* et *VS* sont exprimés en millions. Le score de Dawid-Sebastiani est divisé par dix.
Source : voir premiers paragraphes de la section 3.

diminue lentement au fil du temps, parce que la taille observée de la population se rapproche de l'espérance de taille de la population, ce qui compense l'augmentation de l'écart type de la taille de la population prédite dans les deux prévisions – voir l'expression (3).

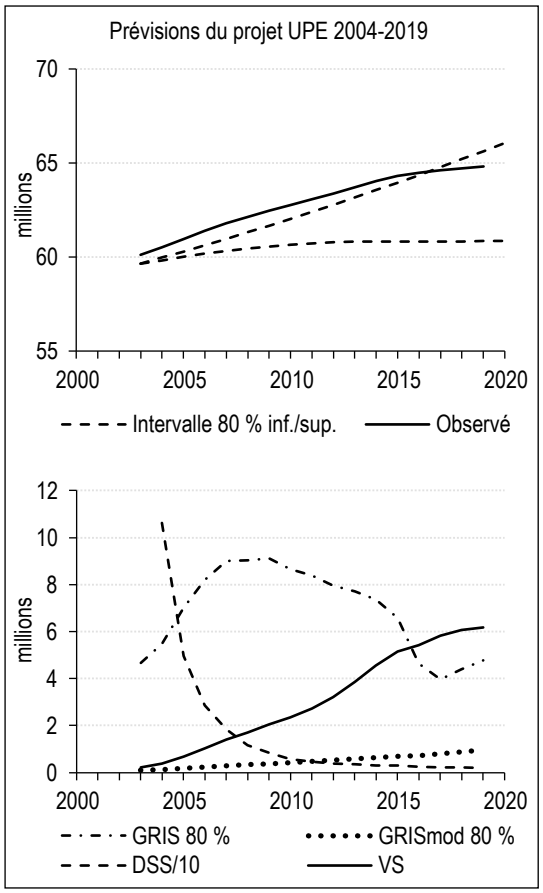
En termes qualitatifs, le *GRIS* affiche la même tendance, plutôt irrégulière, que le *DSS*. C'est particulièrement net pour les Pays-Bas, avec les observations qui sortent d'abord des intervalles mais y reviennent par la suite (figure IV). On constate des irrégularités semblables (mais d'une ampleur beaucoup moins importante) pour la Norvège (cf. figure III). De plus, le *GRISmod* et le *VS* évoluent de façon homogène aux Pays-Bas, comme nous l'avons vu pour la Norvège.

La figure V montre les scores du projet UPE pour la taille de la population totale de la France métropolitaine. Caractéristique frappante : la prévision de population de base, en 2003, est inférieure de près de 500 000 personnes à l'estimation actuelle de la taille de la population pour cette même année. Des données fournies par Eurostat, mises à disposition en 2004, sont à la base des simulations du projet UPE. Les valeurs observées de la figure V proviennent de l'Insee (voir https://www.insee.fr/en/statistiques/serie/000067670). Il va de soi que les chiffres de la population de 2003 fournis par Eurostat en 2004 ont été révisés par la suite.

L'erreur sur la population de base entraîne de très mauvaises valeurs pour les fonctions de notation Gneiting-Raftery (non modifiée) et Dawid-Sebastiani. Quels auraient été ces scores si les prévisions du projet UPE avaient démarré avec l'estimation révisée de la taille de la population totale pour 2003 (60.102 millions) plutôt qu'au niveau effectivement utilisé (59.635 millions) ? Nous pouvons donner une réponse approximative[5] en augmentant de 467 000 personnes l'intervalle de prédiction de 80 %. La figure VI montre les résultats, selon les mêmes échelles verticales que celles de la figure V. Le *DSS* s'améliore considérablement, passant à 5.2 en 2005 et à 5.6 en 2006 (contre

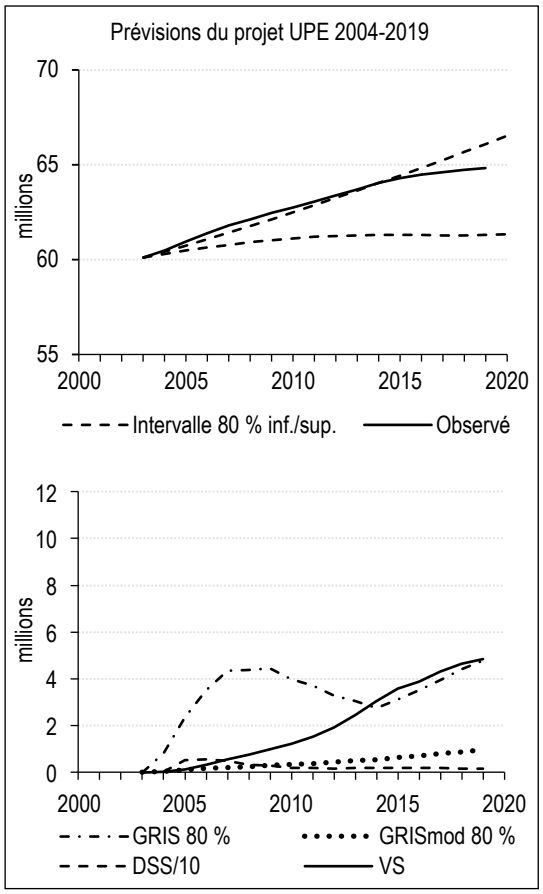
5. Approximative parce que nous ne tenons pas compte des conséquences d'une population de base plus importante en termes de fécondité et de mortalité.

Figure V – **Taille de la population totale, France métropolitaine. Intervalles de prédiction, valeurs observées et scores**



Note : les intervalles de prédiction et les valeurs observées sont présentés dans la partie supérieure, tandis que les scores d'intervalle (*GRIS* et *GRISmod*), le score de Dawid-Sebastiani (*DSS*) et le score basé sur la variance (*VS*) sont présentés dans la partie inférieure. Les intervalles de prédiction, les valeurs observées et les scores *GRIS, GRISmod* et *VS* sont exprimés en millions. Le score de Dawid-Sebastiani est divisé par dix.
Source : voir premiers paragraphes de la section 3.

Figure VI – **Taille de la population totale, France métropolitaine. Intervalles de prédiction, valeurs observées et scores, base 2003 révisée**



Note : les intervalles de prédiction et les valeurs observées sont présentés dans la partie supérieure, tandis que les scores d'intervalle (*GRIS* et *GRISmod*), le score de Dawid-Sebastiani (*DSS*) et le score basé sur la variance (*VS*) sont présentés dans la partie inférieure. Les intervalles de prédiction des prévisions du projet UPE 2004-2019 sont corrigés de l'erreur inhérente à la population de base. Les intervalles de prédiction, les valeurs observées et les scores *GRIS, GRISmod* et *VS* sont exprimés en millions. Le score de Dawid-Sebastiani est divisé par dix.
Source : voir premiers paragraphes de la section 3.

respectivement 49.6 et 28.6), puis se stabilise aux environs de 1.6 ou 1.7 après 2015 (au lieu de redescendre doucement vers 2.0 en 2019). Les scores d'intervalle et le score basé sur la variance diminuent légèrement. Ces résultats soulignent combien il est important de choisir la bonne population de base. Les chiffres de la population sont souvent révisés, notamment dans les pays qui ne disposent pas de registre de la population. Dans ce cas, il convient de traiter la population de base comme étant stochastique, en plus des paramètres de fécondité, de mortalité et de migration. Alho & Spencer (2005) donnent un exemple de valeurs de base aléatoires pour une prévision de population probabiliste applicable à la Lituanie.

Un constat commun qui ressort à ce stade est que plus nous projetons loin dans le futur, plus le *GRISmod* et le *VS* se détériorent parce que les intervalles de prédiction s'élargissent et les variances des distributions prédictives augmentent. Bien sûr, cela reflète le fait que les prévisions de population sont plus difficiles à établir sur le long terme que sur le court terme. Contrairement au *GRISmod* et au *VS*, le *DSS* se stabilise à mesure que les délais de réalisation des prévisions augmentent. L'explication se trouve dans la définition de cette fonction de notation spécifique, qui est la somme des deux termes : un terme augmente tandis que l'autre diminue lorsque la variance de la prédiction augmente – voir l'expression (3). En conséquence, nous pouvons dire que le *DSS* n'est pas une mesure appropriée pour analyser la rapidité avec laquelle la qualité d'une prévision se détériore lorsque le délai de réalisation augmente. Toutefois, nous pouvons également dire que le *DSS* permet de

contrôler les effets du délai de réalisation d'une prévision, précisément parce qu'il change très peu au fil du temps. Une autre possibilité consiste à examiner la pente des courbes du *GRISmod* et du *VS*, puisque ces deux fonctions de notation augmentent de façon plutôt homogène avec le temps. Des recherches complémentaires à ce sujet, s'appuyant sur les données de nombreuses autres prévisions (et tenant compte de la taille de différentes populations – voir ci-dessous), sont indispensables.

Comme nous l'avons dit, les scores relativement mauvais de la France peuvent s'expliquer par le fait que les fonctions de notation dépendent de la taille de la population. Pour pouvoir comparer les différents pays, nous devons normaliser les scores. Le tableau 1 montre les résultats des cinq prévisions en 2018, ces scores étant normalisés (voir section 2.4.).

Les scores de la prévision française et des deux prévisions néerlandaises de l'année 2018 sont alors très semblables – voir la partie supérieure du tableau 1. Dans de nombreux cas, les scores des deux pays sont d'un ordre de grandeur supérieur à ceux de la Norvège. Pendant de nombreuses années, la taille de la population observée en France et aux Pays-Bas est restée dans les limites des intervalles de prédiction (voir la partie supérieure des figures IV et VI, sachant que les intervalles français sont corrigés de l'erreur sur la population de base), ce qui contribue aux bons scores des deux pays.

Les deux prévisions relatives à la Norvège reçoivent un mauvais score en raison de la sous-prédiction de l'immigration nette, comme nous l'avons indiqué plus haut. Une raison supplémentaire expliquant les scores élevés du projet StocProj en 2018 est que l'année de base de cette prévision est 1996. Le délai de réalisation, pour atteindre 2018, est de 22 ans, donc plus long que celui du projet UPE (15 ans pour atteindre 2018). La partie inférieure du tableau 1 montre les scores normalisés du projet StocProj en 2011, donc après 15 ans. Par rapport aux scores des deux autres pays au bout de 15 ans, la situation s'est fortement améliorée, mais les scores du projet StocProj restent beaucoup plus élevés que ceux de CBS et du projet UPE en France et aux Pays-Bas.

L'évaluation finale des prévisions de la taille de la population totale se fait au moyen du *CRPS*. Nous le calculons à l'aide de 3 000 simulations tirées du projet UPE pour 2010. Le *CRPS* dépend de la taille de la population – voir l'expression (6). Pour optimiser les comparaisons entre les trois pays, le tableau 2 présente les scores normalisés, définis comme étant le *CRPS* divisé par la moyenne des 3 000 simulations. Les résultats confirment la bonne qualité des prévisions du projet UPE pour les Pays-Bas.

3.2. Structures par âge et par sexe

Les figures VII à IX donnent les *CRPS* normalisés des populations simulées, par sexe et par tranches d'âge de cinq années, au 1er janvier 2010, selon les prévisions du projet UPE. Les lignes horizontales en pointillé représentent les valeurs du *CRPS* pour les tailles de la population totale du tableau 2. Les trois graphiques utilisent la même échelle verticale.

Les tendances des scores par âge varient fortement d'un pays à l'autre. Les résultats pour la Norvège à la figure VII sont faciles à interpréter. Les notes élevées, c'est-à-dire les prévisions de moindre qualité, concernent les jeunes enfants, les jeunes adultes et les personnes âgées. Les scores sont bien meilleurs dans les tranches 10-19 ans et 55-74 ans. Cette tendance par âge

Tableau 1 – Scores d'intervalle, score basé sur la variance et score de Dawid-Sebastiani

	Norvège		Pays-Bas		France[a]
	StocProj	UPE	CBS	UPE	UPE
Année 2018					
GRIS/μ	0.564	0.513	0.062	0.053	0.069
GRISmod/μ	0.038	0.022	0.021	0.011	0.014
VS/μ^2 (x 1000)	17.552	6.569	1.108	0.781	1.154
DSS – 2ln(μ)	-1.525	2.073	-6.797	-7.149	-6.639
15 ans dans le futur					
GRIS/μ	0.231	0.513	0.049	0.053	0.069
GRISmod/μ	0.021	0.022	0.016	0.011	0.014
VS/μ^2 (x 1000)	4.870	6.569	0.906	0.781	1.154
DSS – 2ln(μ)	-3.752	2.073	-6.903	-7.149	-6.639

[a] Chiffres corrigés de l'erreur sur la population de base.
Note : tous les scores sont normalisés.
Source : voir premiers paragraphes de la section 3.

Tableau 2 – *CRPS* normalisés pour la taille de la population totale, prévisions UPE pour 2010

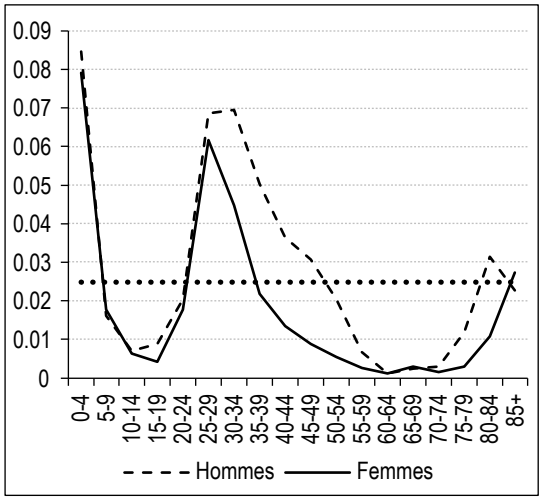
Norvège	Pays-Bas	France
0.0249	0.0075	0.0492

Source : voir premiers paragraphes de la section 3.

reflète la sous-prédiction de l'immigration après 2005, comme nous l'avons noté plus haut, mais les erreurs de prédiction relatives aux naissances et aux décès ont pu également jouer un rôle. De fait, la tendance des scores par âge est similaire, en termes qualitatifs, à celle des erreurs absolues des prévisions ponctuelles de la structure par âge et par sexe des pays industrialisés (cf. par exemple Keilman, 2009). Cela vient du fait que les naissances, les flux migratoires et les décès sont difficiles à prédire. Le délai de réalisation des prévisions du projet UPE n'est que de sept ans. Avec un horizon si court, la fécondité n'a pas d'impact sur la tranche d'âge 10-19 ans. La migration internationale et la mortalité n'influencent que très peu ces tranches d'âge. Il en est de même pour la tranche d'âge 55-74 ans. Il va de soi que, si l'évaluation avait eu lieu après un délai de réalisation de vingt ans ou plus, les valeurs du *CRPS* normalisé des tranches d'âge 10-19 ans et 55-74 ans auraient été bien plus mauvaises. Pour finir, notons que les scores attribués aux hommes dans les tranches d'âge 19-54 ans et 75 ans et plus sont un peu supérieurs à ceux attribués aux femmes des mêmes tranches d'âge. En effet, les hommes sont plus susceptibles que les femmes de migrer (entre 19 et 54 ans) ou de mourir (après 75 ans).

Alors que le score de la Norvège correspond à ce que l'on pouvait attendre, ceux des deux autres pays sont plus difficiles à interpréter. Les scores normalisés indiquent que la prévision néerlandaise est de meilleure qualité que les deux autres, à l'exception de la tranche des personnes âgées. Les scores français tendent à diminuer à mesure que l'âge augmente. Cette tendance suggère que la fécondité était plus difficile à prédire avec exactitude que la migration internationale ou la mortalité. Nous pourrions également avancer plusieurs autres explications. Tout d'abord, la révision des chiffres de la population pourrait avoir été plus prononcée dans certaines tranches d'âge que dans d'autres. Nous avons constaté (chiffres non fournis ici) que les chiffres révisés des hommes et des femmes, par tranche d'âge de cinq années, sont supérieurs d'environ 1 % à ceux utilisés dans le projet UPE. Soulignons cependant quelques exceptions. Les révisions se chiffrent à moins de 0.5 % dans les tranches d'âge 0-4 ans et 80 ans et plus, tandis que, pour les hommes de 20-24 ans, le chiffre révisé est inférieur de 1 % à celui utilisé dans le projet UPE. Cette tendance née des révisions effectuées entre 2003 et 2010 n'est pas illustrée dans la figure IX. Ensuite, la surestimation ou la sous-estimation des flux migratoires nets vers la France entre 2003 et 2009 peut également varier d'une tranche d'âge à l'autre. Enfin, nos données empiriques sur la distribution par âge et par sexe, à partir de 2010, incluent les effets de corrections administratives, terme qui couvre la correction d'erreurs d'enregistrement et des ajustements statistiques. Ces corrections sont nécessaires au cas où l'enregistrement des naissances et des décès est incomplet. En Norvège et aux Pays-Bas (pays disposant d'un registre de la population), les erreurs d'enregistrement de l'immigration et de l'émigration sont également incluses dans les corrections administratives. Les effets de ces corrections sont probablement minimes en Norvège ; ils sont plus marqués pour les Pays-Bas et la France. Par exemple, les données de CBS et de l'Insee montrent que la migration nette totale de la période 2003-2009, sans correction, se chiffre à 214 000 personnes aux Pays-Bas et 601 000 en France. Mais avec les données de migration nette d'Eurostat, qui incluent ces corrections, les chiffres sont nettement différents sur la même période, à savoir respectivement 17 000 et 884 000 personnes[6]. Dans la mesure où il est difficile d'obtenir des données fiables dissociant la migration nette et les corrections administratives et ajustements par tranche d'âge aux Pays-Bas et en France, nous

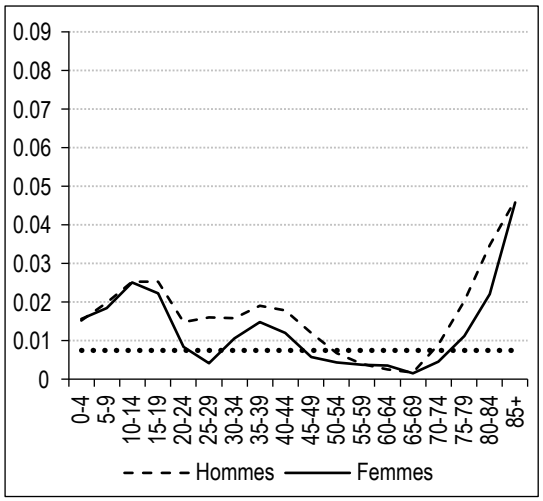
Figure VII – ***CRPS* normalisés pour la population par âge et sexe, Norvège, prévisions du projet UPE, 2010**



Source : voir premiers paragraphes de la section 3.

6. En Norvège, les chiffres sont de 188 300 (sans correction) et de 187 800 (avec corrections). Pour la France, l'Insee fournit des chiffres distincts pour la migration nette et les ajustements. Ce détail n'est pas disponible dans les données d'Eurostat.

Figure VIII – ***CRPS* normalisés pour la population par âge et par sexe, Pays-Bas, prévisions du projet UPE, 2010**



Source : voir premiers paragraphes de la section 3.

Figure IX – ***CRPS* normalisés pour la population par âge et par sexe, France métropolitaine, prévisions du projet UPE, 2010**

Source : voir premiers paragraphes de la section 3.

n'avons pas analysé ce problème de façon plus poussée. Notons également que les prévisions du projet UPE ne contiennent pas de variable séparée qui traite des corrections administratives (pratique courante pour les prévisions de population).

Nous tirons de cette évaluation la conclusion générale suivante : les prévisions du projet UPE concernant la pyramide de la population néerlandaise de 2010, telle que mesurée par le *CRPS* normalisé, sont de meilleure qualité que les prévisions du projet UPE concernant la Norvège et la France, à l'exception des personnes très âgées. La tendance par âge du *CRPS* de la Norvège est semblable à celle des erreurs absolues des prévisions ponctuelles. Il est difficile de dire pourquoi les tendances par âge varient grandement entre les trois pays, notamment en raison de problèmes liés aux données sur la migration internationale.

* *
*

Cet article vise à montrer comment une prévision de population probabiliste peut être évaluée, une fois que les observations relatives aux variables prédites sont disponibles. Les statisticiens ont développé diverses règles de notation à ces fins, mais elles sont très peu appliquées dans la littérature relative aux prévisions de population. Une règle de notation mesure la distance entre la distribution de probabilité de la variable prédite et ses résultats réels. En soi, un score n'a pas de signification intrinsèque – nous ne pouvons l'interpréter qu'en le comparant au score d'une autre prévision. Nous avons utilisé les règles de notation qui récompensent l'exactitude (le résultat est proche de l'espérance de la prédiction) et la précision (la distribution prédictive présente une faible variance, de sorte qu'il est difficile d'atteindre l'objectif). On peut arguer que l'exactitude est plus importante que la précision : la précision ne devrait être récompensée que si le résultat n'est pas trop loin de la tendance centrale de la distribution prédictive. Nous avons discuté la notion de « trop loin ».

Un prévisionniste peut mettre ses prévisions probabilistes à la disposition des utilisateurs de trois façons différentes. Premièrement, il peut publier un intervalle de prédiction pour la variable d'intérêt. Des probabilités de couverture de 67 % et de 80 % sont les plus courantes. Certains prévisionnistes traitant de la population présentent des intervalles de prédiction de 95 %. Nous recommandons d'éviter cette pratique, car les intervalles de 95 % sont très larges, dans la mesure où ils tendent vers les quantiles en cas de survenance d'événements extrêmes. Deuxièmement, le prévisionniste peut fournir à l'utilisateur une base de données contenant les parcours d'échantillonnage de l'évolution de la taille de la population, simulée de façon stochastique, ainsi que des résultats d'autres prévisions. Parfois, seuls le moment de premier ordre (espérance) et le moment de second ordre (variance) de l'intervalle de prédiction sont disponibles. Nous avons présenté les règles de notation pouvant être utilisées pour l'un ou l'autre des types de résultats de prévisions. Les règles de notation sont orientées négativement : un score inférieur indique une meilleure prévision.

Nous avons évalué les prévisions de population probabilistes établies pour la France, les Pays-Bas et la Norvège. Pour les trois pays, nous avons utilisé les résultats du projet UPE. Puisque de nombreuses règles de notation appliquent la même échelle que pour la taille de la population, nous avons proposé d'utiliser des règles de notation normalisées lorsque l'intérêt réside dans la comparaison des prévisions établies pour différents pays. Nous avons examiné les intervalles de prédiction relatifs à la taille de la population sur la période 2004-2019, ainsi que 3 000 parcours d'échantillonnage relatifs aux pyramides des âges pour l'année 2010. Aux Pays-Bas et en Norvège, nous avons comparé les résultats du projet UPE avec les conclusions des prévisions de population probabilistes officielles du CBS (2001-2019) et d'une prévision probabiliste concernant la Norvège (1997-2019). Toutes les prévisions ont été calculées à l'aide de la méthode par cohorte et composante et selon des paramètres stochastiquement variables pour la fécondité, la mortalité et la migration, ainsi qu'une population de base déterministe.

Nos évaluations montrent que les prévisions du projet UPE concernant les Pays-Bas et la Norvège ont obtenu de meilleurs scores que les autres prévisions pour ces deux pays, parce que les prédictions du projet UPE étaient relativement précises, avec des intervalles de prédiction étroits. Les prévisions du projet UPE concernant la France sont basées sur la population de base de 2003, estimée à 60.1 millions de personnes au moment où la prévision a été calculée. Ce chiffre dépasse de près de 500 000 personnes l'estimation actuelle de la population de 2003 (59.6 millions). L'erreur sur la population de base a engendré un mauvais score pour la prévision française. Il est courant de réviser les statistiques sur la population des années intercensitaires une fois que les données tirées d'un nouveau recensement sont disponibles. Si l'on n'est pas certain de la taille et de la structure d'une population durant une période intercensitaire, la bonne approche consiste à traiter la population de base de la prévision comme étant stochastique.

Nous avons évalué les 3 000 simulations du projet UPE relatives à la composition par âge et par sexe prédite pour 2010. Une fois normalisées en fonction des chiffres de la population pour chaque tranche d'âge et chaque sexe, les prédictions relatives aux Pays-Bas ont reçu les meilleurs scores, à l'exception de la tranche d'âge des personnes très âgées. Pour le score norvégien, la tendance par âge reflète la sous-prédiction de l'immigration après l'élargissement de l'Union européenne en 2005. Toutefois, les erreurs de prédiction relatives à la fécondité et à la mortalité ont pu elles aussi jouer un rôle. Les scores de la France spécifiques à chaque tranche d'âge sont difficiles à interpréter. Ils ne reflètent pas la tendance par âge de la révision susmentionnée des données sur la population de 2003. La sur-prédiction ou la sous-prédiction de la fécondité, de la mortalité et de la migration ont pu elles aussi jouer un rôle. Dans le modèle par cohortes et composantes, la composition de la population de 2010 par âge et par sexe est une fonction non-linéaire complexe des paramètres du modèle relatifs à la mortalité, à la fécondité et à la migration avant 2010. Pour cette raison, nous ne pouvons pas identifier l'impact de ces trois composantes de changement sur les scores.

En plus du problème de la révision des données, nous avons également été confrontés à celui des « corrections administratives ». Ces corrections sont parfois utilisées par les instituts de statistique en tant que composante distincte de changement de la structure et de la taille de la population. En cas d'erreur dans l'enregistrement des naissances, des décès et des flux migratoires, il est nécessaire de faire ces corrections administratives et des ajustements statistiques pour obtenir des statistiques de la population cohérentes en termes comptables. Ces corrections influencent fortement les chiffres empiriques de la population pour les Pays-Bas et la France.

La littérature évaluant les prévisions de probabilité et examinant de nombreuses règles de notation est abondante. Un grand nombre de ces règles s'appliquent à la distribution prédictive d'une variable aléatoire discrète et présentent peu d'intérêt pour l'évaluation des prévisions démographiques. Si nous nous limitons aux règles de notation applicables aux variables aléatoires continues, la littérature en recense également un grand nombre et nous n'en avons choisi que quelques-unes. Comme nous l'avons montré aux sections 2 et 3, ces règles de notation sont très différentes, accordant par exemple des poids différents à la distance ou à la précision. Certaines règles attribuent un mauvais score dès que les chiffres observés sortent de l'intervalle de prédiction. D'autres évoluent de façon plus homogène à mesure que l'observation s'éloigne de la tendance centrale et des bornes de l'intervalle. Des travaux supplémentaires sont nécessaires sur les règles de notation applicables aux prévisions démographiques probabilistes, qui, nous l'espérons, permettront de définir des principes directeurs guidant la sélection de ces règles dans diverses situations.

Les règles de notation sont utiles dans le cadre de l'évaluation *ex-post facto* de deux prévisions

probabilistes ou plus. Après avoir déterminé qu'une prévision est meilleure qu'une autre, sur la base de plusieurs fonctions de notation, nous devons nous demander pourquoi. Pour répondre à cette question, il faut soigneusement analyser les nombreuses étapes du processus de production des deux prévisions probabilistes. Cela constitue un nouveau défi, notamment lorsque des spécialistes appartenant à des institutions différentes ont calculé les deux prévisions. □

BIBLIOGRAPHIE

Alders, M. & De Beer, J. (1998). Kansverdeling van de bevolkingsprognose ("Probability distribution of the population forecast"). *Maandstatistiek van de Bevolking*, 46, 8–11.

Alexopoulos, A., Dellaportas, P. & Forster, J.J. (2018). Bayesian forecasting of mortality rates by using latent Gaussian models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2), 689–711. https://doi.org/10.1111/rssa.12422

Alho, J. & Nikander, T. (2004). Uncertain population of Europe: Summary results from a stochastic forecast. http://www.stat.fi/tup/euupe/del12.pdf

Alho, J. & Spencer, B. (2005). *Statistical Demography and Forecasting*. New York: Springer.

Askanazi, R., Diebold, F. X., Schorfheide, F. & Shin, M. (2018). On the comparison of interval forecasts. *Journal of Time Series Analysis*, 39(6), 953–965. https://doi.org/10.1111/jtsa.12426

Bijak, J. & Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies*, 70(1), 1–19. https://doi.org/10.1080/00324728.2015.1122826

Blanpain, N. & Buisson, G. (2016). Projections de population à l'horizon 2070 : Deux fois plus de personnes de 75 ans ou plus qu'en 2013. *Insee Première* N°1619.
https://www.insee.fr/fr/statistiques/fichier/version-html/2496228/ip1619.pdf

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

CBS (2001). Maandstatistiek van de Bevolking N° 49 (januari), pp. 63–70.

Costemalle, V. (2020). Projections probabilistes bayésiennes de population pour la France. *Economie et Statistique / Economics and Statistics*, ce numéro.

Gneiting, T. & Raftery, A. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. https://doi.org/10.1198/016214506000001437

Gneiting, T. & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Applications*, 1, 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

ISTAT – Istituto Nazionale di Statistica (2018). Il futuro demografico del paese: Previsioni regionali della popolazione residente al 2065 (base 1.1.2017). *Report Statistische* 3 maggio 2018. Roma: ISTAT.

Jordan, A., Krüger, F. & Lerch, S. (2019). Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software*, 90(12). https://doi.org/10.18637/jss.v090.i12.

Keilman, N. (1990). *Uncertainty in national population forecasting: Issues, backgrounds, analyses, recommendations*. Amsterdam and Rockland, MA: Swets and Zeitlinger Publishers.

Keilman, N. (2009). Erroneous population forecasts. *In*: P. Festy & J.-P. Sardon (Eds.) *Profession démographe - Hommage à Gérard Calot*, pp. 237–254. Paris: INED.

Keilman, N., Pham, D. Q. & Hetland, A. (2002). Why population forecasts should be probabilistic - illustrated by the case of Norway. *Demographic Research*, 6-15, 409–454.
https://doi.org/10.4054/DemRes.2002.6.15

Keyfitz, N. (1981). The limits of population forecasting. *Population and Development Review*, 8(44), 579–593. https://doi.org/10.2307/1972799

Matheson, J. E. & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096. http://dx.doi.org/10.1287/mnsc.22.10.1087

Murphy, A. (1970). The ranked probability score and the probability score: A comparison. *Monthly Weather Review*, 98(12), 917–924.
https://doi.org/10.1175/1520-0493(1970)098%3C0917:TRPSAT%3E2.3.CO;2

NRC – National Research Council (2000). *Beyond six billion: Forecasting the world's population. Panel on Population Projections*. John Bongaarts and Rudolfo Bulatao (eds). Washington DC: National Academy Press.

Raftery, A., Li, N., Ševčíková, H., Gerland, P. & Heilig, G. (2012). Bayesian probabilistic population projections for all countries. *PNAS - Proceedings of the National Academy of Sciences*, 109(35), 13915–13921. https://doi.org/10.1073/pnas.1211452109

Shang, H. L. (2015). Statistically tested comparisons of the accuracy of forecasting methods for age-specific and sex-specific mortality and life expectancy. *Population Studies*, 69(3), 317–335. https://doi.org/10.1080/00324728.2015.1074268

Shang, H. L., Smith, P., Bijak, J. & Wisniowski, A. (2016). A multilevel functional data method for forecasting population, with an application to the United Kingdom. *International Journal of Forecasting*, 32, 629–649. https://doi.org/10.1016/j.ijforecast.2015.10.002

Shang, H. L. & Hyndman, R. (2017). Grouped functional time series forecasting: An application to age-specific mortality rates. *Journal of Computational and Graphical Statistics*, 26(2), 330–343. https://doi.org/10.1080/10618600.2016.1237877

Smith, S., Tayman, J. & Swanson, D. (2001). *State and Local Population Projections: Methodology and Analysis*. New York: Kluwer Academic/Plenum Publishers.

Staël von Holstein, C.-A. (1970). A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology*, 9, 360–364. https://www.jstor.org/stable/26174966

Statistics New Zealand (2011). National Population Projections: 2011(base) – 2061. Bulletin published 19 July 2012.
http://archive.stats.govt.nz/browse_for_stats/population/estimates_and_projections/NationalPopulationProjections_HOTP2011.aspx (accessed on 21 March 2019).