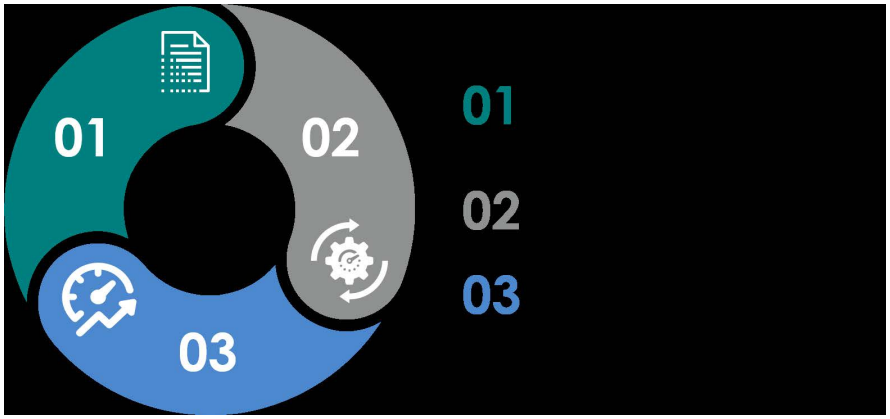
1. L'APPROCHE STATISTIQUE EST AU CŒUR DE LA CONSTITUTION DE MODELES DE PROJECTION

Les modèles statistiques permettent d'utiliser les historiques de mesures de performance d'une activité pour définir des relations mathématiques entre ces mesures et des variables externes (variables macro-économiques, données du marché bancaire, etc.) ou internes (saisonnalité). Les projections consistent alors en des modèles de régression multivariée, linéaires ou non-linéaires[3].

Dans le cadre de ces travaux, un compromis doit souvent être trouvé entre facilité d'implémentation et d'appropriation par les métiers d'une part, et puissance prédictive et robustesse statistique d'autre part. La démarche pour aboutir à ce compromis consiste alors à mettre en œuvre une démarche itérative en trois étapes (comme le détaille la figure 1) :

1. la récupération et la transformation de la donnée,
2. la construction du modèle,
3. l'évaluation de la performance du modèle.

Figure 1 : Processus itératif menant à d'élaboration d'un modèle



01

- **Mises en place organisationnelle permettant de collecter une donnée de qualité.**
- **Transformation pour rendre la donnée exploitable** via la suppression des événements exceptionnels, complétion des données manquantes, dessaisonalisation et autres transformations mathématiques potentielles.
- **Le choix des modèles en étape 2 doit dépendre de la qualité et de la profondeur historique des données.**

02

- **Choix du modèle appliqué –** linéaire en première approche puis utilisation de méthodes de modélisations plus poussées en fonction des données et du type de business étudié.
- Les méthodes d'apprentissage automatique permettent d'orienter les choix des variables, on leur préférera les méthodes plus standards :
  a) Plus faciles à mettre en place
  b) Permet d'obtenir des résultats plus simples à auditer (problématique de la « black-box » des projections en Machine Learning).

03

- Utilisation de **tests statistiques** pour tester la robustesse mathématique des modèles ;
- **Backtesting** des modèles sur des périodes historiques ;
- **Test de la sensibilité** du modèle à un choc sur les variables explicatives.

---

[3] Les modèles de régression permettent d'expliquer l'évolution d'une variable en fonction d'une (modèle univarié) ou plusieurs variables (modèle multivarié). Ces modèles de régression peuvent être linéaires s'il existe une relation de proportionnalité directe entre la variable expliquée et les variables explicatives.

2. LA GESTION DE LA DONNEE : PREREQUIS A LA CONSTRUCTION DE MODELES STABLES

La qualité des approches statistiques dépend fortement des données sur lesquelles s'appuient les travaux de modélisation. En effet, si les données historiques à partir desquelles sont définies les relations avec les indicateurs macro-économiques (ou autres indicateurs) comportent des effets venant les « polluer », les modèles sont moins précis voire peuvent mener à des conclusions erronées. Dans ce contexte, la validation avec les métiers de la qualité des données est un prérequis à toute analyse statistique. Il est important d'impliquer des membres des équipes métiers dans cette réflexion afin de capter l'ensemble des éléments « non-normatifs ». Une fois ces vérifications réalisées, des transformations de séries de données peuvent améliorer la qualité des modèles à tester dans l'étape (2). La pertinence de ces transformations doit tenir compte de différents types d'effets dont :

- la représentativité **d'éléments business dits exceptionnels** (*one-offs*), comme des fusions et acquisitions d'ampleur (« *jumbo deals* ») ou des chocs fiscaux, qui peuvent en réalité être récurrents ;

- la **complétion de données manquantes** et/ou asynchrones. Par exemple, on peut souhaiter projeter mensuellement un indicateur à partir de prédicteurs disponibles seulement trimestriellement. Dans ce cas, il est soit possible de s'appuyer sur des séries trimestrielles (ce qui implique de perdre des points d'analyse), soit d'interpoler mensuellement (de façon linéaire ou non), voire de réaliser un filtrage[4] en complétant les données manquantes par des estimations statistiquement cohérentes issues d'indicateurs ayant une plus grande fréquence ;

- **la saisonnalité[5], à retraiter** par exemple dans le cas de variables explicatives non saisonnières. La transformation de dessaisonalisation peut se baser sur l'algorithme ARIMA-X12[6] ou la procédure STL basée sur des régressions locales[7] ;

- le **lissage de la série à expliquer**, par exemple en calculant une moyenne mobile[8] ;

- **l'introduction d'un effet retard**, ou *lag*, sur la série à expliquer.

D'autres transformations mathématiques pourront alternativement être appliquées sur les séries pour itérativement améliorer les résultats des modèles évalués pendant l'étape d'évaluation des modèles *(3)* :

---

[4] Ceci peut être implémenté à l'aide d'un filtre de Kalman par exemple.

[5] Ou plus généralement l'autocorrélation de la série, ce qui revient à introduire une variable endogène dans le modèle.

[6] L'algorithme ARIMA-X12 est une méthode populaire d'ajustement de la saisonnalité développée par le Bureau du recensement des États-Unis. Cette méthode s'applique à des séries ayant une saisonnalité mensuelle ou trimestrielle. Elle est implémentée dans la plupart des logiciels de statistiques et est l'une des méthodes prônées par l'European Statistical System (ESS).

[7] La procédure STL ("Seasonal and Trend decomposition using Loess") est une méthode de décomposition d'une série temporelle en une composante saisonnière, une tendance et des résidus. A ce titre, c'est aussi une méthode d'ajustement de la saisonnalité qui peut être préférée dans certains cas à des méthodes du types ARIMA-X12 (notamment en cas de composante saisonnière fluctuante ou en présence d'observations aberrantes).

[8] Plus généralement, cela peut être incorporé, avec la saisonnalité, dans une démarché de modélisation de type ARMA (Auto Regressive Moving Average), ARIMA (AutoRegressive Integrated Moving Average) ou SARIMA (Seasonal ARIMA).

- **la différentiation des données** – d'ordre un ou deux – avec une fréquence courte (un trimestre par exemple) ou longue (une année par exemple). Ceci permet généralement de corriger le biais statistique de la non-stationnarité[9], mais peut parfois se révéler instable en projection ;

- l'application de fonctions (rendement, carré, logarithme) visant à **capter des effets non linéaires** ;

- **l'utilisation de modèles en co-intégration** (voir ci-dessous) en cas de non-stationnarité des variables à expliquer.

L'appréciation de la qualité des données collectées ainsi que les caractéristiques des transformations qui leur ont été appliquées pendant l'étape de collecte de la donnée *(1)* doivent être prises en compte dans le choix du modèle développé pendant l'étape de modélisation *(2)*. En effet, la profondeur historique des données doit être suffisante pour capter des situations distinctes (crises, scénarios de taux différenciés, etc.) et les données doivent présenter des réalités métier comparables (par exemple un effet d'échelle lié au nombre de traders sur un desk ou une réorganisation de l'activité doivent être pris en compte pour harmoniser les séries d'un point de vue statistique).

3. ADAPTER LA DETERMINATION DES VARIABLES EXPLICATIVES ET L'EXPRESSION DU MODELE A L'ENVIRONNEMENT

Choix du modèle

L'expression du modèle doit s'adapter aux besoins des utilisateurs de l'outil.

- Les approches linéaires, par exemple, sont plus simples à mettre en œuvre mais ne permettent pas de capter des relations plus complexes qu'une relation affine entre la variable (ou sa croissance) et son inducteur (ou sa croissance). Couplées à l'utilisation de transformations non linéaires sur les variables explicatives du modèle, les approches linéaires simples permettent cependant de capter des non-linéarités. Par exemple, le logarithme des crédits immobiliers peut être corrélé au logarithme de la croissance du revenu des ménages. L'utilisation de la transformation logarithmique permet d'associer des variables dont l'ordre de grandeur est différent.

- Les méthodes d'apprentissage automatique[10],comme les forêts aléatoires[11], sont de très bons outils pour orienter le choix des variables. Elles sont par contre rarement conservées pour la projection elle-même car elles sont souvent complexes à mettre en place et difficiles à auditer pour le régulateur. En outre, elles ne mettent pas toujours en exergue de *driver* exogène de l'activité et peuvent rester trop centrées sur des formes autorégressives[12].

---

[9] Le caractère stationnaire (ou non) d'une série temporelle se rapporte à l'homogénéité de sa distribution statistique au cours du temps. Une propriété plus faible utilisée en pratique (stationnarité faible) est le fait d'avoir ses deux premiers moments (moyenne et variance) constants, ainsi qu'une fonction d'autocorrélation invariante par translation au cours du temps.

[10] Ces méthodes font partie de ce que l'on appelle aujourd'hui le « Machine Learning » qui vise à tirer parti des données pour déterminer la forme du modèle à adopter, plutôt que de le spécifier en amont. Ces méthodes se fondent sur l'analyse statistique d'un grand nombre de données de natures variées.

[11] Les forêts aléatoires sont une famille d'algorithmes d'apprentissage automatique qui se fondent sur des ensembles d'arbres de décision. L'intérêt de cette méthode est d'entrainer un ensemble d'arbres de décision sur des sous-ensembles de l'ensemble des données initiales et ainsi de limiter le problème de sur apprentissage. Ce type d'algorithme permet de faire de la classification (estimation de variables discrètes) mais aussi de la régression (estimation de variables continues).

[12] Un modèle autorégressif est un modèle dans lequel une variable est expliquée par ses valeurs passées plutôt que par d'autres variables.

Modèles en co-intégration

En cas de non-stationnarité[9], les modèles statistiques classiques sont instables et des techniques spécifiques sont à mettre en œuvre. Une notion aujourd'hui centrale est la notion de modèle en *co-intégration* pour des variables macro-économiques. Un ensemble de variables est co-intégré avec la série observée s'il existe une combinaison des variables permettant d'annuler « *la tendance stochastique* » de la série observée pour aboutir à une série stationnaire. Par exemple, il a été démontré qu'aux Etats-Unis, la consommation (réelle, par habitant) et le revenu disponible (réel, par habitant) sont co-intégrés, soulignant une relation stable entre ces deux séries non stationnaires.

Ces variables co-intégrées sont alors liées avec la série observée par une équation linéaire dite « long-terme » qui s'interprète comme un équilibre macro-économique par rapport auquel les écarts constituent des fluctuations temporaires. En reprenant l'exemple précédent, une fluctuation temporaire de la consommation par rapport au revenu disponible peut survenir à un trimestre donné, mais celle-ci se répercutera avec le signe opposé sur la consommation future du prochain trimestre dans une proportion comparable, ce qui tend à ramener les deux séries vers leur point d'équilibre qui est représenté par la relation long-terme.

Les approches historiques pour appréhender ce type de relation sont les approches d'Engle et Granger[13] ou Johannsen[14], ainsi que les modèles dits *Autoregressive Distributed Lag* (ARDL). Tous ces modèles capturent à la fois les relations de long-terme ainsi que les déviations par rapport à ces équilibres via des modèles de retour à la moyenne et de correction d'erreur.

Choix des variables

Dans un premier temps les variables explicatives seront choisies, parmi l'ensemble des variables transformées, grâce à des études de corrélations simples. Le choix des variables explicatives peut aussi être éclairé par l'expertise métier, des approches systématiques de classification (comme une analyse en composantes principales[15]) ou encore leur significativité dans les méthodes d'apprentissage automatique (on conserve alors les variables mises en exergue par la méthode mais en leur appliquant des modèles statistiques classiques).
Au contraire, certaines variables seront exclues a posteriori par les tests statistiques en étape (3). En particulier, un trop grand nombre de variables peut aboutir à du sur-calibrage (*over-fitting*) et des variables colinéaires à une instabilité des coefficients de la régression[16].

Calibrage des paramètres

La méthode d'estimation des paramètres de la régression dépend des tests réalisés en étape (3). Ils seront estimés soit par l'estimateur des moindres carrés, soit par exemple par des estimateurs de Yule-Walker[17] pour éviter les biais que fait peser l'existence d'autocorrélation des résidus dans les séries utilisées.

---

[13] Co-Integration and Error Correction: Representation, Estimation, and Testing, Robert F. Engle and C. W. J. Granger, 1987).

[14] Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models, Johansen, Søren, 1991.

[15] L'analyse en composantes principales (ACP), est une méthode d'analyse de données, qui consiste à transformer des variables corrélées entre elles en nouvelles variables dé-corrélées les unes des autres à partir de leur caractéristiques mathématiques (décomposition orthogonale aux valeurs propres).

[16] Les régressions Lasso ou Ridge permettent de régulariser le problème et de sélectionner les variables de plus d'intérêt en introduisant des termes de pénalité.

[17] Les équations de Yule-Walker établissent une correspondance directe entre les paramètres du modèle et ses autocovariances. Elles sont utiles pour déterminer la fonction d'autocorrélation ou estimer les paramètres d'un modèle.

Les problèmes posés par la non–stationnarité concernent aussi l'inférence des paramètres du modèle estimé, pour laquelle les lois asymptotiques usuelles dérivées dans le cadre de séries stationnaires peuvent mener à des incohérences si utilisées telles quelles. Notamment, les p-values (voir plus bas) et les intervalles de confiances ne sont plus fiables dans le cadre des séries non-stationnaires ou de la co-intégration.

4. L'EVALUATION DES MODELES DONNE SA CREDIBILITE AUX TRAVAUX DE PROJECTION STATISTIQUE

Le pouvoir prédictif du modèle doit être vérifié par un ensemble de tests. Des tests statistiques ou des *backtestings* peuvent être conduits pour conforter le choix d'un modèle, bien qu'aucun ne soit éliminatoire quant au choix du modèle. Notons que l'exigence de vérification de ces tests est à pondérer par la qualité de la donnée disponible. La sensibilité du modèle à un choc des variables explicatives devra dans tous les cas être appréciée.

Tests statistiques

Le calcul de la significativité des variables (*p-value*) est important mais l'estimation des paramètres et le calcul des p-valeurs doivent être corrigés en cas de non-respect des hypothèses de base de la régression linéaire [18] :

- **Stationnarité des séries temporelles**[19] (homogénéité dans le temps de leur distribution) : les résultats des régressions linéaires peuvent être instables dans le temps si les séries ne sont pas stationnaires, même en cas de bon R^2. Dans ce cas, il est préférable de transformer les variables (étape (1)) ou de choisir un modèle en co-intégration (étape (2)).

- **Résidus homoscédastiques**[20] (variance constante dans le temps) et/ou plus généralement non auto-corrélés[21] : en cas de non-respect, cela peut être révélateur d'une autre variable explicative non trouvée. Cela peut biaiser de manière importante les variances et les intervalles de confiance des coefficients. Il est donc nécessaire d'introduire des corrections sur les coefficients[22] ou de modifier les estimateurs utilisés[23].

- **Normalité des résidus**[24] : cette hypothèse de la régression linéaire est cependant rarement vérifiable sur de petits échantillons (propriété asymptotique) et n'est pas nécessaire à la convergence des estimateurs des paramètres.

Backtest (ou cross-validation, ou out-of-sample performance)

Si la profondeur historique l'y autorise, il est possible de mesurer l'écart entre la série historique réelle et le modèle calibré sur une période temporelle différente. En renouvelant l'exercice sur plusieurs sous-périodes, il est possible par ailleurs de vérifier la stabilité des coefficients de la régression. Une erreur moyenne équivalente entre la période testée et la période de calibrage est un bon indicateur que le modèle n'est pas sur-calibré (*over-fitting*).

---

[18] Quand les hypothèses qui fournissent les distributions asymptotiques ou les intervalles de confiance des estimateurs ne sont plus vérifiées, les intervalles de confiance peuvent quand même être calculés par simulation (bootstraping ou ré échantillonnage).

[19] Les tests classiques à mener sont ceux de Dickey-Fuller (augmenté), Phillips-Perron ou Kwiatkowski-Phillips-Schmidt-Shin.

[20] Tests de Breusch-Pagan et de Goldfeld et Quandt.

[21] Tests de Durbin-Watson et de Breusch-Godfrey.

[22] Transformation de Yule-Walker (Cochrane-Orcutt généralisé).

[23] Correction de la matrice de covariance par l'estimateur de Newey-West.

[24] Test de Shapiro-Wilk par exemple.

5. DIFFICULTES

La modélisation de l'ensemble des agrégats financiers de la banque suppose la modélisation d'activités de natures diverses s'appuyant sur des modélisations statistiques hétérogènes. Dans ce contexte, l'équipe en charge du développement des modèles doit s'adapter à la réalité de chacun des segments d'activité. Il en résulte une panoplie de modèles statistiques à articuler sur une plateforme flexible permettant de les lier entre eux et aux sources de données (bases métier notamment) pour proposer des résultats directement exploitables par les équipes de la plateforme.

Quatre types de difficultés doivent être dépassées pour constituer une base de modèles suffisamment solide à intégrer à la plateforme :

- La difficulté à trouver des modèles statistiques prédictifs sur certains périmètres : l'ensemble des activités ne sont pas modélisables par une approche statistique et sont plus complexes à appréhender (commissions spécifiques, frais généraux, éléments exceptionnels etc). Par ailleurs, la plupart des modèles statistiques classiques peinent à capter les non-linéarités dans les comportements passés.

- L'imbrication d'effets distincts à modéliser en parallèle de la modélisation des activités : effets change, concentration des portefeuilles, etc. Les effets de contagion, de réputation et l'ensemble des *feedback effects* sont particulièrement complexes à capter.

- La difficulté à collecter des données de qualité faciles à mettre à jour après le premier exercice de modélisation : manque de profondeur des données, problématiques d'homogénéité etc.

- Les difficultés organisationnelles et d'outil.

6. CONCLUSION

Si les établissements bancaires disposent aujourd'hui d'équipes spécialistes de la modélisation quantitative reconnues, ces compétences se concentrent surtout dans les équipes Risques sur les problématiques de risque de crédit et de risque de marché. Pour la plupart des établissements, la modélisation prospective à partir de méthodes statistiques implique donc la constitution d'équipes spécialistes.

Les méthodes discutées plus haut apportent une vision d'ensemble des moyens statistiques dont disposent les équipes de planification pour construire des modèles de projection. La dimension humaine et la capacité à recruter des talents capables de construire des modèles complexes est au cœur des enjeux.

Le développement d'une approche tactique, via des outils agiles, permet dans un premier temps de constituer un premier support à la plateforme et de distinguer la construction des modèles de leur industrialisation dans les systèmes de la banque.