

Degré de confiance pour les indicateurs de performance : degré de fiabilité du processus de production et écart significatif¹

L'utilisation des indicateurs de performance ne peut se faire de manière pertinente que si les utilisateurs sont avertis du degré de confiance qu'ils peuvent accorder aux résultats. Dans son ouvrage « Performance Indicators for Water Supply Services – Operations & Maintenance Specialist Group », l'International Water Association (IWA) propose une méthode basée sur la détermination d'un degré de fiabilité et d'un niveau de précision.

L'objet du présent document est de présenter une approche similaire basée sur :

1. une méthode de détermination du degré de fiabilité de la production de chaque indicateur, méthode inspirée de celle de l'IWA,
2. la notion d'écart significatif, qui permet de concilier d'une part le besoin de conclure au vu des résultats obtenus et d'autre part d'éviter des conclusions erronées.

Cette nécessaire prudence ne remet pas en cause la pertinence des indicateurs, mais vise à identifier les processus peu fiables d'obtention des indicateurs, afin de conduire le producteur de données à améliorer cette fiabilité, en se posant les bonnes questions, et aboutir à des conclusions plus pertinentes.

En conclusion, les bonnes pratiques et les principes de prudence à retenir dans l'utilisation des résultats des indicateurs de performance sont les suivants :

Pour le producteur de données :

Pour chaque indicateur, le producteur de donnée évalue le degré de fiabilité du processus de production de l'indicateur. Il s'efforce d'améliorer ce processus afin de garantir un certain niveau de rigueur et une traçabilité. Il s'autoévalue chaque année pour en suivre les améliorations.

Pour toute utilisation des résultats :

Par principe, toute comparaison, tout calcul d'écart ne peut se faire que sur des indicateurs pour lesquels **le processus de production des résultats est jugé fiable ou très fiable.**

Si tel est le cas, une analyse ne peut être déclenchée que si les écarts sont significatifs.

Par convention, l'écart significatif est fixé à 2% pour un même service sur plusieurs exercices, et à 5% pour plusieurs services sur un même exercice. Si l'écart est supérieur à ces valeurs, l'analyse des tendances est déclenchée. A l'inverse, on ne peut rien conclure si un indicateur varie de moins de 2% d'une année à l'autre, ou si la différence entre deux services est inférieure à 5%.

La comparaison permet de conclure à des tendances mais ne permet pas de quantifier des écarts de manière définitive.

¹ fiche établie à partir des travaux des producteurs de données (groupe FP2E, FNCCR) et de Ernst & Young sur la base de la méthode développée par l'International Water Association

1 Détermination du degré de fiabilité de la production de chaque indicateur

1.1 Objectif de la démarche

La méthode est mise en place afin d'évaluer la fiabilité de la production des indicateurs de performance. On ne vise pas la fiabilité du résultat en tant que valeur numérique mais **la fiabilité du processus** qui permet d'obtenir ce résultat.

Cette notion de fiabilité, issue des travaux de l'IWA, est utilisée ici pour stimuler les efforts de rigueur et de traçabilité des collectivités et des entreprises, en définissant des niveaux pour positionner les processus de production des indicateurs. Cette approche permet également d'en suivre l'amélioration dans le temps.

Il s'agit d'une méthode d'autoévaluation de la fiabilité, pour permettre à tout service de procéder à sa propre évaluation. Elle est avant tout une incitation à la rigueur et au progrès (un service fera des efforts pour progresser d'une classe de fiabilité).

Cette méthode ne renseigne pas directement sur les résultats. En effet, ce n'est pas le résultat lui-même qui est jugé fiable, mais le processus de production de ce résultat. L'obtention de ce résultat tient compte des écarts liés à la latitude d'interprétation des définitions et du manque de précision des mesures physiques. **L'attribution d'une classe de fiabilité A ou B à un processus de production d'un indicateur ne donne donc pas une indication bien définie sur la précision du résultat** et ne permet pas une utilisation aveugle des chiffres. A contrario, un processus de production d'un indicateur classé C incitera à se méfier encore plus des valeurs obtenues et toute comparaison avec d'autres années ou d'autres services sera sans intérêt, voire dangereuse.

Les tentatives d'interprétation des résultats des indicateurs de performance ne devront se faire que sur des services dont le processus de production de ces indicateurs est de classe A ou B.

Il s'agit d'un préalable avant toute tentative de comparaison entre exercices successifs pour un même service ou entre plusieurs services sur un même exercice. **Aucun calcul d'écarts et aucune analyse sur ces écarts ne devront être conduits sur des services dont le processus de production des indicateurs n'est pas classé A ou B.**

1.2 Principe de détermination du degré de fiabilité

Le producteur de données doit se positionner dans une grille d'évaluation organisée autour de 3 classes de fiabilité et de 4 critères. Les 3 classes de fiabilité sont :

- A pour « très fiable »
- B pour « fiable »
- C pour « peu fiable »

Les critères couvrent l'ensemble du processus de calcul des indicateurs, depuis la mesure sur le terrain ou la prise d'information des données de base jusqu'au résultat du calcul de l'indicateur.

Cette méthode concerne aussi bien les données issues de mesures physiques (telles que les volumes, les analyses de qualité d'eau...) que les autres données (telles que le nombre de réclamations, de débordements...).

Afin de conserver une approche pragmatique et utilisable, l'évaluation de la fiabilité se fait globalement au niveau de l'indicateur, en fonction de la façon dont des données qui rentrent dans sa composition sont collectées et traitées, et non pas au niveau de chaque donnée individuelle.

Pour un indicateur considéré, la démarche est la suivante :

1. le producteur de données se positionne entre les classes A et C pour chacun des critères retenus ;
2. le processus est classé B si tous les critères sont au moins de classe B, il est classé A si tous les critères sont de classe A, il est classé C autrement.

Ainsi, cette méthode permet d'associer à chaque indicateur un degré de fiabilité du processus de production, qui doit être utilisé pour aider à l'interprétation des résultats.

On trouvera aux annexes 1 et 2 le détail des critères et la grille d'autoévaluation avec les 3 classes de fiabilité.

2 Écart significatif

Avoir des processus fiables de production des indicateurs n'est pas suffisant pour s'autoriser à comparer sans discernement les résultats obtenus par différents services ou par un même service sur plusieurs exercices.

En effet, en détaillant le processus d'élaboration d'un indicateur, on voit que les écarts entre les résultats obtenus et la réalité proviennent de la subjectivité des définitions, du manque de fiabilité et du manque de précision :

1. les définitions des indicateurs, après avoir essayé de les préciser pour lever les ambiguïtés, comportent encore une part inévitable d'imprécision et d'interprétation ;
2. le processus de production des indicateurs mis en place par les collectivités et les entreprises est plus ou moins fiable ;
3. le manque de précision des mesures physiques est inhérent à toute mesure.

Par le jeu cumulé de ces écarts, les chiffres se trouvent souvent loin de ce qui est acceptable pour pouvoir en tirer des conclusions sérieuses et utilisables, ce que les retours de terrain viendront certainement démontrer dans les années à venir. Des valeurs d'indicateurs proches ne sont ainsi pas nécessairement significativement différentes et retirent toute pertinence à une comparaison².

Pour un même service, rien ne prouve que l'erreur soit toujours dans le même sens d'une année sur l'autre, et la rigueur conduit à refuser des conclusions bâties sur la persistance d'erreurs systématiques. Quand il s'agit de comparer des services entre eux, les écarts sont potentiellement encore plus importants, et cette rigueur est d'autant plus nécessaire.

Mais tout le monde demande instinctivement des conclusions. En toute rigueur, le processus ne permet pas de répondre naturellement à ces attentes, toutefois on peut définir des règles pour éviter les conclusions trop erronées. Deux solutions sont possibles :

- Soit chercher à déterminer les **vraies caractéristiques de l'incertitude** sur le chiffre obtenu, en analysant tout le processus de production de ce chiffre³. Mais cette entreprise de vérité apparaît trop complexe, et nécessite de plus un consensus scientifique sur tous les détails de calcul. Il paraît ainsi inenvisageable de rechercher cette « vérité » par des calculs d'incertitudes.

² Ainsi, quand on compare un ratio de 79% à un ratio de 81 %, cela ne consiste pas à dire « la comparaison de ces chiffres doit être faite avec prudence, mais quand même, elle veut bien dire quelque chose » : on ne sait absolument pas laquelle des deux situations est supérieure à l'autre, et il faut donc s'interdire de conclure dans un sens ou dans l'autre.

³ Théoriquement, on pourrait trouver qu'il y a 5% de chances qu'un chiffre soit inférieur à 75, 15% de chances qu'il soit entre 75 et 77, 30% de chances qu'il soit entre 77 et 79, etc. On pourrait aussi montrer que le chiffre "81" a 55% de chances d'être réellement supérieur au "79"

- La seule alternative est donc la recherche d'un accord conventionnel : tout le monde sait que les calculs ne sont pas exacts et que les conclusions seront parfois injustes, mais accepte de jouer le jeu par décision conventionnelle. La règle consiste à dire qu'il serait absurde de conclure à un écart significatif au-dessous d'un certain écart conventionnel. Cet écart conventionnel est choisi :
 - inférieur à l'écart significatif réel que l'on peut supposer (qui est souvent supérieur à 20% !);
 - selon un ordre de grandeur psychologique (pas trop grand pour ne pas décourager);
 - le plus possible sécurisant (donc suffisamment grand), sans être inacceptable par des tiers;
 - selon des modalités simples, et de bon sens.

Il est considéré par convention qu'un écart est significatif si les chiffres de deux services diffèrent de plus de 5%, et si deux chiffres d'un même service au cours du temps diffèrent de plus de 2%⁴.

Ce sont les seuils de déclenchement de l'analyse. Si un écart est jugé significatif, on conclura sur une tendance, sans quantifier. Dans tous les cas, chacun se doit de garder à l'esprit l'objectif initial des indicateurs : se mettre en position de réflexion.

⁴ Pour un service, un indicateur prend une valeur X l'année N et Y l'année N+1. Par convention, on considère que l'écart au cours du temps est significatif si $(Y-X)/X$ est supérieur à 2% en valeur absolue.
 Pour deux services, l'année N, un indicateur prend la valeur X dans un service et la valeur Z dans l'autre. Par convention, l'écart est considéré comme significatif si $(Z-X)/Z$ est supérieur à 5% si $Z > X$ ou si $(X-Z)/X$ est supérieur à 5% si $X > Z$.

Annexe 1 : Règle et critères retenus

Les tableaux ci-après précisent :

- La règle permettant d'attribuer un niveau de fiabilité en fonction du niveau de respect des 4 critères
- Pour chaque critère les éléments clés permettant de différencier les classes de fiabilité.

La définition des classes de fiabilité de chaque critère figure en annexe 2.

Tableau 1 – règle d'attribution de la classe de fiabilité de production d'un indicateur

Classe de fiabilité	A	B	C
Règle	100% des critères applicables sont de classe A	100% des critères applicables sont au moins de classe B	un critère (ou plus) applicable est de classe C

Tableau 2 – critères d'évaluation du processus de production d'un indicateur

Critère	Éléments clés (cf. détail par niveau de fiabilité en annexe 2)
1 Procédures et méthodes de calcul	- Existence d'un ensemble cohérent de documents écrits, référencés, accessibles et diffusés décrivant les définitions, les méthodes de calcul ainsi que les rôles et responsabilités en matière de collecte, de calcul et contrôles (notion de procédure)
2 Traçabilité	- Existence d'une base de données de référence ou de supports papiers partagés et accessibles
3 Contrôles et validation	- Contrôles des données - Validation de l'indicateur par l'encadrement
4 Metrologie (mesures physiques), le cas échéant	- Suivi des meilleures pratiques (a) - Importance des estimations effectuées (seuil d'un tiers)

(a) suivi des meilleures pratiques (métrologie) : mesure validée par les autorités de contrôles, ou conforme aux prescriptions réglementaires lorsqu'il en existe (ex : compteurs d'eau froide), ou réalisée par un laboratoire accrédité lorsque ce dispositif existe. Pour les autres cas, les mesures sont effectuées par du personnel qualifié et habilité, suivant une procédure écrite et avec du matériel dont les performances métrologiques sont périodiquement vérifiées.

Pour les indicateurs ne faisant intervenir aucune donnée issue de mesures physiques, le critère n°4 (métrologie) ne s'applique pas.

L'autoévaluation doit être représentative de l'ensemble de l'exercice auquel sont rattachées les valeurs des indicateurs de performance. Les modifications significatives effectuées en cours d'exercice susceptibles de modifier l'autoévaluation seront prises en compte lorsqu'elles seront rétroactivement appliquées sur la quasi-totalité de l'exercice (ex : définition ou contrôle des indicateurs).

Note sur l'articulation de l'évaluation entre l'indicateur et les données le composant :

La fiabilité s'apprécie de manière globale sur le processus d'élaboration de l'indicateur. Toutefois, lorsque l'indicateur est calculé à partir de données provenant de processus distincts (exemple numérateur et dénominateur de l'indice linéaire des volumes non comptés), le niveau de fiabilité associé à chaque critère (ex : contrôles, procédures, etc.) est le plus petit des niveaux obtenus par chaque processus de production de données. Exemple, pour l'indice linéaire des volumes non comptés, pour chacun des critères, le niveau de fiabilité sera le plus petit du niveau de fiabilité obtenu entre le processus de mesure des volumes non comptés et le processus de mesure des longueurs de réseaux.

Annexe 2 : Grille d'autoévaluation du degré de fiabilité de la production d'un indicateur

La grille d'autoévaluation détaillée dans le tableau ci-dessous repose sur 4 critères d'évaluation et 3 classes de fiabilité.

Critère/ classe	A	B	C
1 Procédures et méthodes de calcul	Il existe un ensemble cohérent de documents écrits, référencés, accessibles et diffusés décrivant les définitions (définition de l'indicateur et de chacune des données qui contribue à son calcul), les méthodes de calcul ainsi que les rôles et responsabilités en matière de collecte, de calcul et de contrôles (notion de procédure)	Il existe des documents écrits décrivant les définitions, les méthodes de calcul ainsi que les rôles et responsabilités en matière de collecte, de calcul et de contrôles sans être systématiquement cohérents, référencés, accessibles et diffusés (ex : courriel, note de service, compte rendu, ...)	Les documents ne décrivent pas l'ensemble des définitions, méthodes de calcul et responsabilités (ou autre)
2 Traçabilité	L'indicateur et les données sont chacun tracés dans une base de données de référence du service, servant à toutes les utilisations et accessibles à plusieurs personnes	L'indicateur et toutes les données sont tracés sur des supports référencés (papiers ou base de donnée). Certains supports ne peuvent être accessibles qu'à une seule personne.	L'indicateur et les données ne sont pas tous tracés sur un support de référence (ou autres cas)
3 Contrôles et validation	L'indicateur est validé formellement à minima annuellement par une personne de l'encadrement. Les données sont enregistrées et contrôlées dans un délai raisonnable (sous un mois pour des activités quotidiennes ou avant la campagne suivante pour des activités périodiques) à compter du constat de l'événement (ex : PV de réception ou d'analyse). Le contrôle peut consister en des tests automatiques ou manuels effectués par une personne (tests de vraisemblance, analyses statistiques, etc.).	L'indicateur est validé formellement annuellement par une personne de l'encadrement. Les données sont contrôlées lors du calcul de l'indicateur, par des tests automatiques ou par une personne (test de vraisemblance, analyses statistiques, etc.).	L'indicateur n'est pas formellement validé par l'encadrement ou les données ne font pas l'objet de contrôles lors de leur acquisition ou du calcul de l'indicateur (ou autres cas)
4 Métrologie (le cas échéant)	Les mesures suivent les meilleures pratiques (a) et ne font pas l'objet d'estimation significative (moins de 5 % du total annuel de la donnée considérée).	Les mesures suivent les meilleures pratiques (a). Les estimations sont supérieures à 5% mais restent inférieures à un tiers du total annuel de la donnée considérée.	Absence de suivi des meilleures pratiques (a) ou estimations très significatives (supérieures à un tiers du total annuel de la donnée considérée)

(a) suivi des meilleures pratiques (métrologie) : mesure validée par les autorités de contrôles, ou conforme aux prescriptions réglementaires lorsqu'il en existe (ex : compteurs d'eau froide), ou réalisée par un laboratoire accrédité lorsque ce dispositif existe. Pour les autres cas, les mesures sont effectuées par du personnel qualifié et habilité, suivant une procédure écrite et avec du matériel dont les performances métrologiques sont périodiquement vérifiées.

Pour les indicateurs ne faisant intervenir aucune donnée issue de mesures physiques, le critère n°4 (métrologie) ne s'applique pas.