Unsupervised Learning – Kreditvergabe österreichischer Banken anhand Einzelkreditdaten entschlüsseln

Lorenz Riess, Johannes Temme, Andreas Wolf[1]

1 Einleitung

Österreichische Kreditinstitute sind verpflichtet, Einzeldaten zu Krediten oberhalb gewisser Meldegrenzen[2] in der Granularen Kreditdatenerhebung (GKE) an die OeNB zu melden. Zu jedem Kreditinstrument steht eine Vielzahl an Informationen zur Verfügung, die u. a. das Kreditvolumen, die gültigen Zinssätze, die vorhandenen Sicherheiten und einen regulatorischen Kreditrisikoparameter (Ausfallswahrscheinlichkeit) beinhalten. Diese Datenmengen erlauben einen mikroskopisch genauen Blick auf das Kreditgeschäft der Kreditinstitute und können genutzt werden, um das Kreditgeschäft mehrerer Banken gegenüberzustellen.

In diesem Bericht wird eine Methode vorgestellt, wie Banken aufgrund ihrer gemeldeten Einzelkredite verglichen werden können, ohne den hohen Informationsinhalt der Daten durch bloß einfache Aggregierungsfunktionen (Summe, Mittelwert, Median etc.) außer Acht zu lassen. Bildlich gesprochen erlaubt dieser paarweise Vergleich zwischen Banken, eine Bankenlandschaft zu erstellen. Entscheidend hierfür ist es, Kredite einer Bank als Wahrscheinlichkeitsverteilung aufzufassen, die das Kreditportfolio der Bank darstellt. Dieser Bericht stellt lediglich die Theorie und die verwendete Methodik vor und umfasst keine Analyse anhand konkreter Meldedaten.

Im Folgenden wird daher zuerst erläutert, wie Einzelkreditinformationen einer Bank als Wahrscheinlichkeitsverteilung aufgefasst werden können (Kapitel 2). Anschließend wird diskutiert, wie Unterschiede zwischen den Kreditportfolios von Banken anhand der Abstände ihrer Wahrscheinlichkeitsverteilungen mithilfe der sogenannten „Wasserstein-Distanzen“ berechnet werden können (Kapitel 3). Die Theorien der Wasserstein-Distanzen und des sogenannten „Optimalen Transports“ erlauben es auch, Durchschnitte der Kreditportfolios der Banken zu bestimmen (Kapitel 4). Sobald Durchschnitte und Abstände zwischen Kreditportfolios berechnet werden können, können Banken anhand dieser mit Methoden des „Unsupervised Learning“ gruppiert (geclustert) werden (Kapitel 5). Dies ermöglicht es, in Kapitel 6 eine Bankenlandschaft in Österreich zu visualisieren. Banken mit einer ähnlichen Kreditvergabe werden dabei nahe beieinander dargestellt, während Banken mit sehr unterschiedlicher Vergabepraxis weit voneinander entfernt visualisiert werden. Eine solche Darstellung als Bankenlandschaft ermöglicht es, die Kreditvergabe der Banken anhand ihrer einzelnen Kredite zu entschlüsseln. Kapitel 7 fasst die Erkenntnisse zusammen.

[1] *Oesterreichische Nationalbank, Abteilung Statistik – Abteilung Statistik – Aufsicht, Modelle und Bonitätsanalysen, lorenz.riess@oenb.at, johannes.temme@oenb.at, andreas.wolf@oenb.at.*

[2] *Sämtliche Kredite von Rechtsträgern iSd Art. 1(5) der AnaCredit-VO mit einem Engagement über 25.000 EUR sowie alle Kredite natürlicher Personen mit einem Engagement über 350.000 EUR: siehe Hirsch, Kemetmüller und Lingo, 2020.*

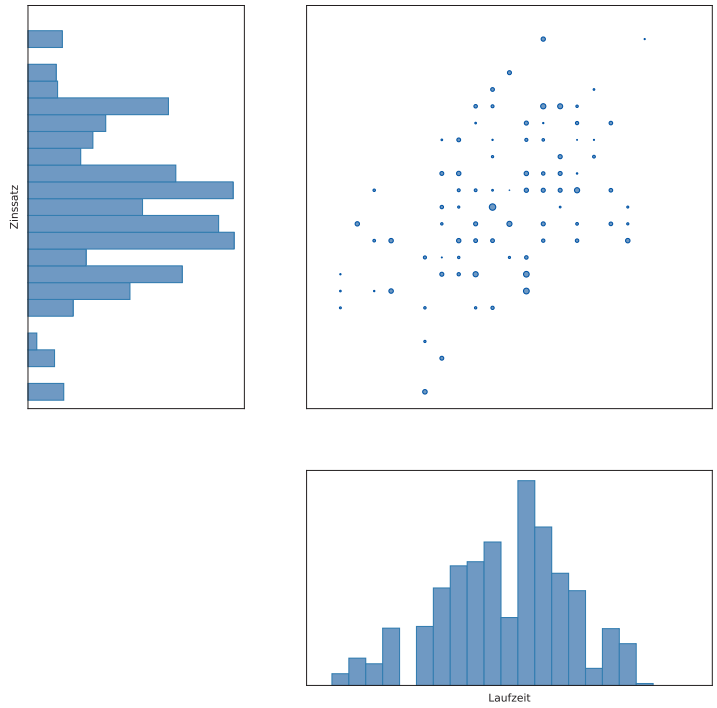
2 Granulare Kreditdaten als Wahrscheinlichkeitsverteilung des Kreditportfolios einer Bank

Zur Illustration nehmen wir anfangs an, dass für eine gegebene Bank an einem gegebenen Stichtag zu jedem Kredit das Kreditvolumen, die Laufzeit und der vereinbarte Jahreszinssatz gemeldet werden. Einen solchen Kredit kann man sich dann als Punkt in der (zweidimensionalen) Ebene aus Zinssatz und Laufzeit vorstellen, wobei die Größe des Punkts[3] vom Anteil des Kredits am gesamthaften Kreditvolumen der Bank bestimmt ist. Eine Bank meldet aber nicht nur einen einzelnen Kredit, sondern entsprechend ihres Kreditportfolios eine Vielzahl von Krediten. Diese gesamte Meldung der Einzelkredite ergibt dann eine Punktwolke in der Ebene aus Laufzeit und Jahreszinssatz (siehe Abbildung 1).

Eine Punktwolke kann man mathematisch als diskrete Wahrscheinlichkeitsverteilung interpretieren. In dem vereinfachten Beispiel aus Abbildung 1 entspricht die Meldung der Bank einer Wahrscheinlichkeitsverteilung im zweidimensionalen Raum. Diese Wahrscheinlichkeitsverteilung beschreibt die vergebenen Kredite einer Bank zum gegebenen Meldestichtag anhand der gewählten Attribute Laufzeit, Jahreszinssatz und Kreditvolumen. Intuitiv und für die folgende Erklärung hilfreich kann man sich die diskrete Wahrscheinlichkeitsverteilung als Kombination von Sandstapeln vorstellen: die Punkte geben an, wo im Raum die einzelnen Sandstapel liegen, während die Höhen der Stapel und somit die Anzahl der verwendeten Sandkörner den Gewichten der Sandstapel entsprechen und die Bedeutung der Punkte angeben, die in Abbildung 1 im Streudiagramm über die Größe der Punkte dargestellt wurde.[4] Wie bei Wahrscheinlichkeitsverteilungen üblich, summieren sich die Höhen der Sandstapel auf 1 (100 %). Die Histogramme in Abbildung 1 zeigen gerade die Sandstapel der eindimensionalen Randverteilungen, beispielsweise im unteren Teil der Abbildung nur die Wahrscheinlichkeitsverteilung der Laufzeiten der Kredite. Analog zu Sandstapeln kann man sich eine Wahrscheinlichkeitsverteilung

Abbildung 1

Punktwolke der einzelnen gemeldeten Kredite einer Bank zu den Attributen Zinssatz und Laufzeit



Quelle: OeNB.

Anmerkung: Die Größe der Punkte entspricht dem Kreditvolumen des jeweiligen Kredits. Links und unterhalb der Punktwolke werden die jeweils eindimensionalen Randverteilungen bzgl. Zinssatz und Laufzeit dargestellt.

[3] *Abweichend vom geometrischen Konzept eines Punktes bezeichnen wir (unterschiedlich große) Kreise in den Abbildungen als Punkte.*

[4] *In Kapitel 3 wird die Idee der Sandstapel und Sandkörner verwendet, um Sandkörner zwischen zwei Verteilungen „optimal" zu transportieren und Verteilungen somit zu vergleichen. In der Informatik ist diese Analogie auch als Earth-Mover-Distanz bekannt.*

auch als Löcher in einem Sandboden vorstellen, wobei die Tiefe der Löcher wieder die Bedeutung der Punkte angibt. Diese Analogie wird in Abbildung 2 verwendet.

Tatsächlich werden weitaus mehr als die im obigen Illustrationsbeispiel genannten drei Attribute je Kreditinstrument gemeldet (siehe Bachmann et al., 2021). Die Analogie zwischen Punkten und Wahrscheinlichkeitsverteilungen bleibt jedoch bestehen, auch wenn sie sich nicht mehr visualisieren lässt wie in Abbildung 1: Wenn neben dem Kreditvolumen weitere d Attribute herangezogen werden, kann jedes Kreditinstrument als Vektor $x \in \mathbb{R}^d$ im d-dimensionalen Raum aufgefasst werden. Die Meldung aller Kredite einer Bank entspricht somit erneut einer Punktwolke im d-dimensionalen Raum und kann daher als Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}^d$ aufgefasst werden. Konkret können wir

$$\mu = \sum_{i=1}^{n} p_i \delta_{x_i}$$

schreiben, wobei $x_1, \dots, x_n \in \mathbb{R}^d$ die Attributsvektoren der gemeldeten Kredite sind und $p_1, \dots, p_n \geq 0, \sum_{i=1}^{n} p_i = 1$ Gewichte beschreiben, die jeder Kredit in der Wahrscheinlichkeitsverteilung hat. Wie im Illustrationsbeispiel oben können wir den Anteil eines Kredits am Gesamtkreditvolumen als Gewicht p_i nehmen. In der Formel beschreibt δ_{x_i} eine sogenannte Punktverteilung (Dirac-Maß) am Punkt x_i der gemeldeten d Attribute des i-ten Kredits. In der obigen Illustration entspricht eine Punktverteilung gerade einem Sandstapel. Die Verteilung μ der Punktwolke ergibt sich somit aus dem gewichteten Mittel von Punktverteilungen δ_{x_i}.

Diese Überlegungen zeigen, dass die vergebenen Kredite einer Bank immer als Wahrscheinlichkeitsverteilung der gewählten Kreditattribute aufgefasst werden können. Wie oben angemerkt ist zu beachten, dass die Meldedaten der GKE nur Kredite oberhalb gewisser Meldegrenzen beinhalten. Die Daten der GKE liefern uns daher zwar meist ein sehr gutes, aber nicht notwendigerweise vollständiges Bild der vergebenen Kredite einer Bank.

Da das Kreditportfolio jeder Bank zu einem Stichtag einer Wahrscheinlichkeitsverteilung der vergebenen Kredite entspricht, stellt sich die Frage, wie man die Kreditportfolios einzelner Banken anhand dieser Wahrscheinlichkeitsverteilungen vergleichen kann. Dafür ist es notwendig, einen Abstand zwischen Wahrscheinlichkeitsverteilungen messen zu können. Banken, deren vergebene Kredite anhand der gewählten Attribute (z. B. Zinssatz und Laufzeit) ähnlich sind, sollten dabei einen geringen Abstand aufweisen, während der Abstand zwischen Banken, deren Kredite auf Basis der Attribute sehr unterschiedlich sind, größer sein sollte. Zu beachten ist dabei, dass beim Vergleich der Wahrscheinlichkeitsverteilungen der Banken nur die Anteile der jeweiligen Kredite am Gesamtforderungswert und nicht etwa die einzelnen absoluten Forderungswerte entscheidend sind. Das folgende Kapitel 3 erläutert, wie ein solcher Abstand zwischen zwei Wahrscheinlichkeitsverteilungen anhand der sogenannten „Wasserstein-Distanz“ berechnet werden kann.

3 Wasserstein-Distanz: Abstände zwischen Banken als Wahrscheinlichkeitsverteilungen berechnen

Wir betrachten im Folgenden zwei Wahrscheinlichkeitsverteilungen μ, ν an vergebenen Krediten zweier Banken. Zunächst wird der Einfachheit halber angenommen,

dass jede Wahrscheinlichkeitsverteilung nur aus je einem Kredit $x \in \mathbb{R}^d$ bzw. $y \in \mathbb{R}^d$ besteht. Wie oben beschrieben, stellt jeder dieser Kredite einen Punkt im d-dimensionalen Raum $\mathbb{R}^d$ dar und ein natürlicher Abstand zwischen den zwei Krediten ist beispielsweise der sogenannte euklidische Abstand

$$\|x - y\|_2 = \left(\sum_{j=1}^{d} (x_j - y_j)^2 \right)^{\frac{1}{2}},$$

der die Länge des Verbindungsvektors zwischen x und y misst. Der Abstand der zwei Wahrscheinlichkeitsverteilungen μ,ν kann daher in diesem Spezialfall mit $\|x - y\|_2$ gleichgesetzt werden.

Im Allgemeinen bestehen μ, ν jedoch aus mehreren und unterschiedlich vielen Krediten $x_1, \dots, x_n \in \mathbb{R}^d$, $y_1, \dots, y_m \in \mathbb{R}^d$, die jeweils andere Gewichte $(p_i)_{i=1}^n, (q_j)_{j=1}^m$ haben, d. h. $\mu = \sum_{i=1}^n p_i \delta_{x_i}, \nu = \sum_{j=1}^m q_j \delta_{y_j}$. Ein naiver Ansatz wäre es, als Abstand zwischen μ und ν genau $\left(\sum_{i=1}^n \sum_{j=1}^m p_i q_j \|x_i - y_j\|_2^2\right)^{\frac{1}{2}}$ zu nehmen. Dies entspricht der gewichteten Summe aller möglichen Abstände zwischen den Krediten von μ und ν, wobei als Gewicht jeweils das Produkt der Gewichte zweier Kredite verwendet wird, um den Abstand zwischen den Wahrscheinlichkeitsverteilungen zu messen. Diese Variante eines Abstands liefert jedoch kein gewünschtes Ergebnis, da jeder Kredit von μ mit jedem Kredit von ν verglichen wird.[5] Es erscheint intuitiv sinnvoller, dass ein Kredit von μ möglichst nur mit jenen Krediten von ν verglichen werden soll, die dem Kredit von μ ähnlich sind, im Idealfall sogar nur mit dem ähnlichsten Kredit. Das Forschungsgebiet der optimalen Transporttheorie liefert hierfür eine passende Antwort, um eine Distanz zwischen Wahrscheinlichkeitsverteilungen, die sogenannte Wasserstein-Distanz, zu berechnen.[6] Die Wasserstein-Distanz $W_2(\mu,\nu)$ kann als Lösung eines sogenannten optimalen Transportproblems definiert werden

$$W_2(\mu,\nu) := \left(\min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \|x_i - y_j\|_2^2 \right)^{\frac{1}{2}},$$

wobei $\Pi(\mu,\nu)$ die Menge aller Matrizen $\pi = (\pi_{ij})_{\substack{i=1\dots n \\ j=1\dots m}} \geq 0$ ist, deren Zeilensummen die Gewichte $(p_1,\dots p_n)$ von μ und deren Spaltensummen die Gewichte $(q_1,\dots q_m)$ von ν sind. Im Vergleich zum naiven Ansatz $\left(\sum_{i=1}^n \sum_{j=1}^m p_i q_j \|x_i - y_j\|_2^2\right)^{\frac{1}{2}}$ kann π eine beliebige andere Gewichtung der paarweisen Abstände zwischen $x_1, \dots, x_n \in \mathbb{R}^d$ und $x_1, \dots, x_n \in \mathbb{R}^d$ sein, solange jeder Punkt x_i insgesamt sein ursprüngliches Gewicht p_i und jeder Punkt y_j sein ursprüngliches Gewicht q_j bekommt. Die Wasserstein-Distanz verwendet nun die „beste" Gewichtung, das heißt jene Gewichtung, welche den geringsten Wert liefert.

Es kann gezeigt werden, dass $W_2(\mu,\nu)$ gerade die minimalen Kosten beim Transport von Punktmassen zwischen μ und ν liefert, wobei die Kosten c beim Transportieren eines Punkts $x \in \mathbb{R}^d$ zu einem Punkt $y \in \mathbb{R}^d$ durch $c(x,y) = \|x - y\|_2^2$ berechnet werden. Um bei der obigen Illustration einer diskreten Wahrscheinlichkeitsverteilung als Sandstapel bzw. -löcher zu bleiben, gibt $W_2(\mu,\nu)$ gerade die

[5] *Insbesondere erfüllt dieser Ansatz nicht alle mathematischen Anforderungen einer Distanz. Beispielsweise liefert dieser Ansatz nicht null, wenn man den Abstand einer Wahrscheinlichkeitsverteilung zu sich selbst berechnet.*

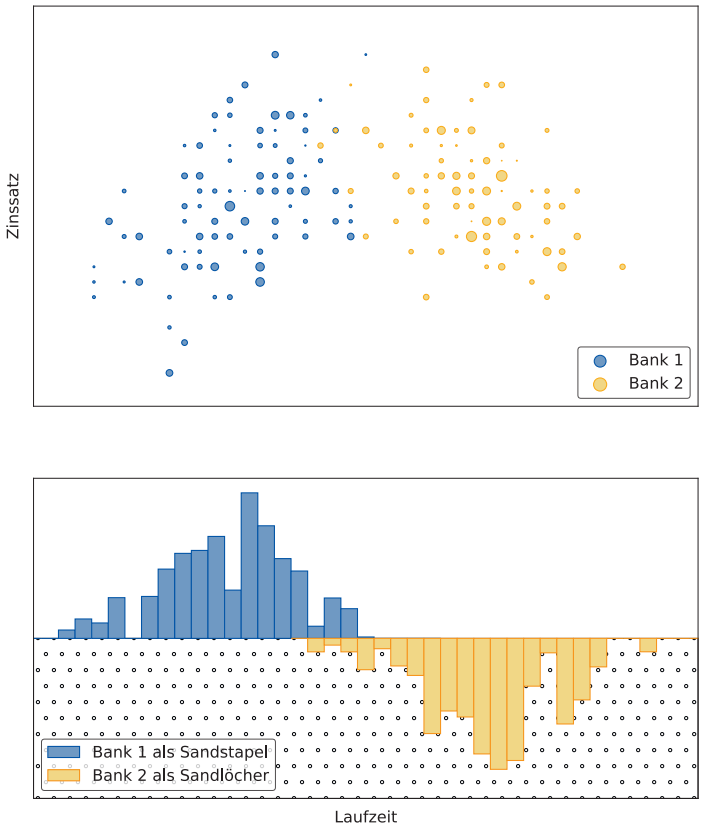
[6] *Die Wasserstein-Distanz ist nach dem russisch-amerikanischen Mathematiker Leonid Nison Vaserštein benannt.*

geringsten Kosten an, um einen Sandstapel μ in die Sandlöcher ν (und vice versa) zu verschieben. Dafür kann es notwendig sein, große Sandstapel auf mehrere Sandlöcher aufzuteilen bzw. mehrere kleine Sandstapel großen Sandlöchern zuzuweisen. Insbesondere erfüllt die Wasserstein-Distanz die oben genannten Anforderungen: wenn sich zwei Wahrscheinlichkeitsverteilungen weniger oder mehr (anhand der Kredite und Gewichte) unterscheiden, wird der Abstand zwischen den Wahrscheinlichkeitsverteilungen kleiner oder größer ausfallen.

Um die Wasserstein-Distanz darzustellen, zeigt der obere Teil von Abbildung 2 die Kreditverteilungen zweier Banken anhand der Attribute Jahreszinssatz und Laufzeit. Man erkennt, dass die Wahrscheinlichkeitsverteilung von Bank 2 (visualisiert durch gelbe Punkte) weiter rechts ist. Die obige erwähnte Intuition der Wasserstein-Distanz als optimaler Transport zwischen Sandstapeln und Sandlöchern ist im unteren Teil von Abbildung 2 dargestellt. Die Wahrscheinlichkeitsverteilung der Laufzeit der Kredite von Bank 1 ist durch blaue Sandstapel visualisiert, während die Wahrscheinlichkeitsverteilung der Laufzeit der Kredite von Bank 2 durch gelbe Sandlöcher dargestellt wird. Wird nun die Frage gestellt, welche Sandkörner der Sandstapel in welche Sandlöcher transportiert werden müssen, sodass insgesamt der gesamte Sand in den Sandlöchern landet, diese voll ausfüllt, und insgesamt der Sand am wenigsten bewegt werden muss, so ist die Antwort darauf genau der optimale Transport, welcher den Wert der Wasserstein-Distanz zwischen den zwei Wahrscheinlichkeitsverteilungen liefert. Diesen Transport kann man tatsächlich als kontinuierliche Bewegung der Punkte der linken Wahrscheinlichkeitsverteilung zu Punkten der rechten Wahrscheinlichkeitsverteilung und somit als „Film“ darstellen, der als Ausgangsbild die linke Wahrscheinlichkeitsverteilung und als Endbild die rechte Wahrscheinlichkeitsverteilung zeigt. Stoppt man den Film genau in der Mitte, erhält man eine Art Mittelwert zwischen den beiden Wahrscheinlichkeitsverteilungen, das sogenannte Wasserstein-Baryzentrum (siehe Agueh und Carlier, 2011). In Kapitel 4 wird dieses genauer beschrieben und in Abbildung 5 anhand des bereits betrachteten Beispiels der zwei Banken als Punktwolken veranschaulicht.

Abbildung 2

Kreditverteilungen zweier Banken



Quelle: OeNB.

Anmerkung: Der obere Teil der Abbildung zeigt die Kreditverteilungen zweier Banken anhand der Attribute Laufzeit und Jahreszinssatz; der untere Teil die eindimensionalen Randverteilungen der Laufzeit der Kredite der Banken (für Bank 1 dargestellt als Sandstapel, für Bank 2 dargestellt als Sandlöcher).

4 Optimale Transporttheorie: Durchschnitte von Banken als Wahrscheinlichkeitsverteilungen berechnen

Im vorherigen Kapitel wurde erläutert, wie sich ein Abstand zwischen zwei Banken aufgrund ihrer Kreditportfolios mit Hilfe der Wasserstein-Distanz ermitteln lässt. Zusätzlich wurde das Wasserstein-Baryzentrum als eine Art Mittelwert zwischen zwei Banken bereits erwähnt. Im Folgenden wird genauer beschrieben, wie ein Wasserstein-Baryzentrum als „Durchschnitt" von zwei oder mehreren Banken gefunden werden kann. Mathematisch stellt sich die Frage, wie ein sinnvoller Durchschnitt von Wahrscheinlichkeitsverteilungen gebildet werden kann. Auch in diesem Fall liefert die optimale Transporttheorie eine Antwort.

Im euklidischen Raum $\mathbb{R}^d$ ist der (arithmetische) Durchschnitt von $x_1,\dots,x_n \in \mathbb{R}^d$ definiert als $\frac{1}{n}\Sigma_{i=1}^{n} x_i$. Er löst aber ebenso eindeutig folgendes Varianz-Minimierungs-Problem, nämlich

$$min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \|x_i - x\|_2^2.$$

Intuitiv ist ein Vektor x gesucht, der möglichst nahe an allen Punkten „gemeinsam" ist. Ganz analog wurde von (Agueh & Carlier, 2011) das Wasserstein-Baryzentrum definiert, wobei die Punkte durch Wahrscheinlichkeitsverteilungen und die euklidische Distanz durch die Wasserstein-Distanz ersetzt werden. Für Wahrscheinlichkeitsverteilungen $\mu_1,\dots,\mu_n$ ist ihr Wasserstein-Baryzentrum also als Lösung des folgenden „Varianz"-Minimierungs-Problems definiert:

$$min_{\nu} \sum_{i=1}^{n} W_2^2(\mu_i, \nu).$$

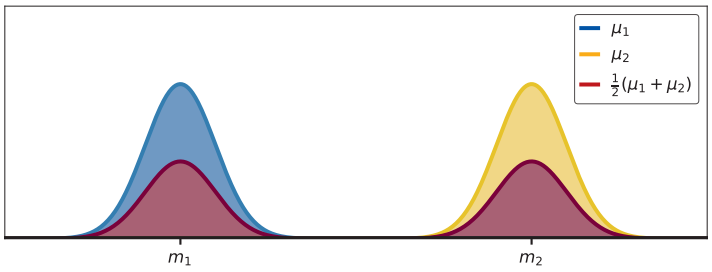
Analog zum euklidischen Fall wird nach jener Wahrscheinlichkeitsverteilung gesucht, die möglichst nah an den betrachteten Wahrscheinlichkeitsverteilungen liegt, wobei der Abstand mit der Wasserstein-Distanz gemessen wird. Das Wasserstein-Baryzentrum eröffnet die Möglichkeit, aus einer gegebenen Anzahl an Wahrscheinlichkeitsverteilungen von Banken eine Durchschnittsbank zu berechnen. Insbesondere kann man so das Kreditportfolio einer durchschnittlichen österreichischen Bank als Wahrscheinlichkeitsverteilung darstellen.

Es gibt für den Durchschnitt von Wahrscheinlichkeitsverteilungen auch andere Möglichkeiten als das Wasserstein-Baryzentrum. Beispielsweise könnte die Konvexkombination von Wahrscheinlichkeitsverteilungen verwendet werden, da diese wieder eine Wahrscheinlichkeitsverteilung ergibt. Es ist jedoch der Fall, dass das Wasserstein-Baryzentrum intuitiver ist und mathematisch ansprechendere Ergebnisse liefert. Um eine Intuition zu geben, betrachten wir zwei Normalverteilungen $\mu_1=N(m_1,\sigma_1^2),\mu_2=N(m_2,\sigma_2^2)$ mit gleicher Varianz, aber unterschiedlichen Mittelwerten. Wird als Durchschnitt die Konvexkombination $\frac{\mu_1+\mu_2}{2}$ betrachtet, so ergibt dies eine Wahrscheinlichkeitsverteilung, die keine Normalverteilung mehr ist (siehe Abbildung 3). Das Wasserstein-Baryzentrum zwischen μ_1 und μ_2 ist hingegen $\mathcal{N}\left(\frac{m_1+m_2}{2}, \sigma^2\right)$, also die Normalverteilung mit gleicher Varianz und als Mittelwert der arithmetische Durchschnitt der Mittelwerte von μ_1 und μ_2 (siehe Abbildung 4).

Zusätzlich zu diesem visuellen Beispiel mit Normalverteilungen, kann auch das Wasserstein-Baryzentrum der zwei Banken, die bereits in Kapitel 3 und in Abbildung 2 betrachtet wurden, visualisiert werden. Dieses ist im unteren Teil von Abbildung 5 dargestellt, während oben nochmals die Kreditverteilungen der zwei fiktiven Banken dargestellt sind. Betrachtet man die Kreditverteilung von Bank 1, lässt sich erkennen, dass es einen positiven Zusammenhang zwischen der Laufzeit und dem Zinssatz eines Kredits gibt. Es ist sichtbar, dass ein Kredit (d. h. ein blauer Punkt) mit höherer Laufzeit eher einen höheren Zinssatz hat. Insbesondere ist die blaue Punktwolke nach „rechts oben" orientiert. Für die gelbe Punktwolke ist genau das Gegenteil der Fall, nämlich scheint es, als würden Kredite, die von Bank 2 vergeben werden, bei höherer Laufzeit eher geringere Zinssätze aufweisen. Beim Wasserstein-Baryzentrum ist erkennbar, dass es diesen Zusammenhang nicht gibt. Dies erscheint intuitiv, da es einmal einen positiven und einmal einen negativen Zusammenhang gibt, und diese Zusammenhänge ähnlich stark erscheinen, ein Durchschnitt also keinen Zusammenhang aufweisen sollte. Zusätzlich ist in der grünen Punktwolke des Wasserstein-Baryzentrums erkennbar, dass die Punkte auch generell „in der Mitte" der blauen und gelben Punktwolke liegen. Auch „Lücken" der blauen und gelben Punktwolken lassen sich im Wasserstein-Baryzentrum erkennen. Beispielsweise weisen sowohl die blaue als

Abbildung 3

Zwei Normalverteilungen (blau, gelb) mit gleicher Varianz und unterschiedlichen Mittelwerten

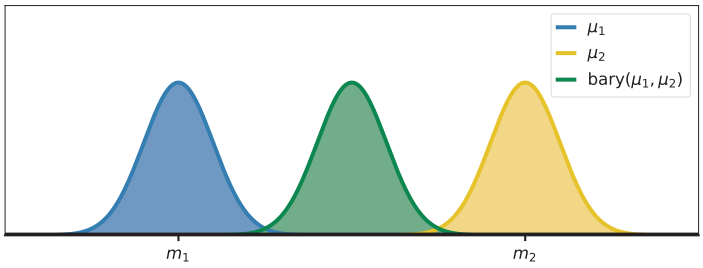


Quelle: OeNB.

Anmerkung: In Rot ist die Konvexkombination der beiden Wahrscheinlichkeitsverteilungen dargestellt, welche als eine Art Durchschnitt der zwei Wahrscheinlichkeitsverteilungen gesehen werden kann, jedoch keine Normalverteilung mehr ist.

Abbildung 4

Zwei Normalverteilungen (blau, gelb) mit gleicher Varianz und unterschiedlichen Mittelwerten

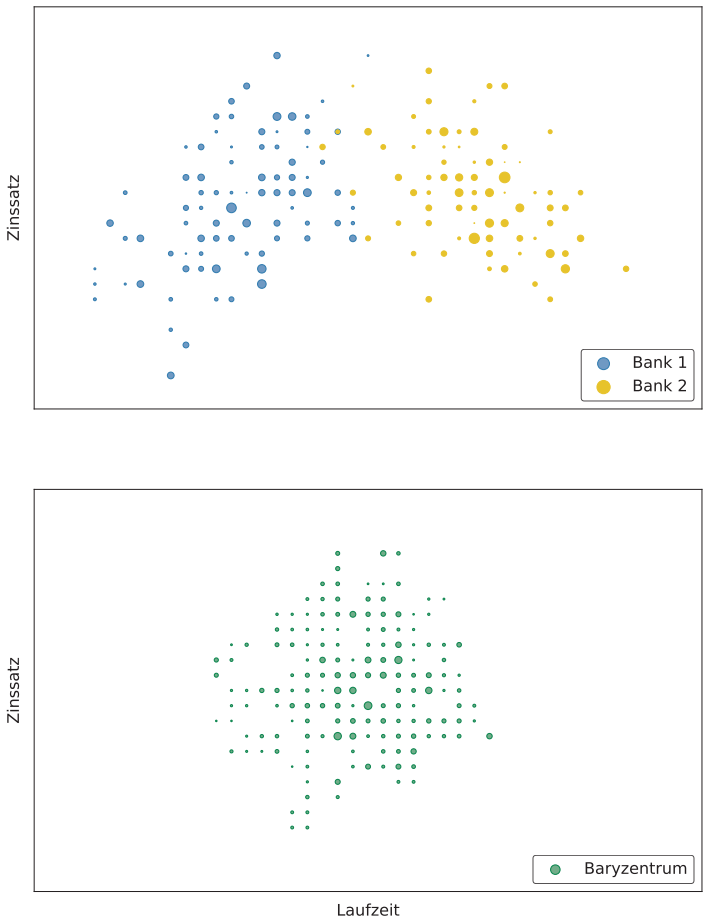


Quelle: OeNB.

Anmerkung: In Grün ist das Baryzentrum der beiden Wahrscheinlichkeitsverteilungen dargestellt, welches auch als Durchschnitt der zwei Wahrscheinlichkeitsverteilungen gesehen werden kann. Dies ist eine Normalverteilung mit der gleichen Varianz, welche als Mittelwert den Durchschnitt der Mittelwerte der anderen Wahrscheinlichkeitsverteilungen hat.

Abbildung 5

Zwei Normalverteilungen (blau, gelb) mit gleicher Varianz und unterschiedlichen Mittelwerten



Quelle: OeNB.

Anmerkung: (oben): Der obere Teil der Abbildung zeigt Punktwolken der gemeldeten Kredite zweier Banken zu den Attributen Zinssatz und Laufzeit, wobei die Größe der Punkte dem Kreditvolumen der jeweiligen Kredite entspricht. Der untere Teil der Abbildung zeigt das Baryzentrum der beiden Punktwolken, das als Durchschnittsbank der beiden obigen Banken interpretiert werden kann, wenn die Punktwolken als Wahrscheinlichkeitsverteilungen betrachtet werden.

die gelbe Punktwolke links eine „Lücke" auf, welche sich auch links in der grünen Punktwolke wieder findet. Insgesamt lässt sich das Wasserstein-Baryzentrum also als eine Durchschnittsbank auf Basis der Banken 1 und 2 interpretieren.

5 Clustering: Gruppieren von Banken

Sobald man Abstände zwischen den Wahrscheinlichkeitsverteilungen der Banken und Durchschnitte bzw. Baryzentren solcher berechnen kann, ist es möglich, ähnliche Banken in einer sogenannten Cluster-Analyse anhand ihrer Wahrscheinlichkeitsverteilungen zu gruppieren. *k*-Means ist dafür ein bekannter Algorithmus, der eine Menge von Objekten in eine vordefinierte Anzahl von *k* Gruppen unterteilt, sodass die Objekte jeder Gruppe möglichst ähnlich sind. Die drei Schritte des Algorithmus werden im Folgenden allgemein vorgestellt:

1. Man definiert die Anzahl *k* der zu findenden Gruppen (Cluster) und wählt zufällig *k* viele Cluster-Baryzentren aus.
2. Für jede Bank misst man den Abstand zu den *k* Cluster-Baryzentren und bestimmt das ihr nächste Cluster-Baryzentrum (mit dem geringsten Abstand). Man weist dadurch jede Bank einem Cluster zu.
3. Innerhalb jedes Clusters bestimmt man das Baryzentrum aus den dem Cluster zugewiesenen Banken.

Im Algorithmus werden die letzten zwei Schritte wiederholt, bis ein Abbruchkriterium erfüllt ist, z. B. bis sich die Zuweisung der Banken zu Clustern nicht mehr ändert. Eine solche Cluster-Analyse ist ein Verfahren des sogenannten „Unsupervised Learning", in dem in den Daten nach vorher unbekannten Mustern (Ähnlichkeiten) gesucht wird, um die Datenpunkte zu gruppieren. Im Gegensatz dazu muss im sogenannten „Supervised Learning" die richtige Gruppierung bereits beim Finden der Muster bekannt sein.

Der Algorithmus funktioniert grundsätzlich mit jeder Methode zum Berechnen eines Abstands, insbesondere auch für Wasserstein-Distanzen. Somit liefert er eine Möglichkeit, Banken anhand der von ihnen vergebenen Kredite zu gruppieren. Bei einer geeigneten Anzahl *k* der zu findenden Cluster sollten Banken mit einem

ähnlichen Kreditportfolio möglichst dem gleichen Cluster zugeteilt werden, während Banken mit sehr unterschiedlichen Krediten in jeweils anderen Clustern landen sollten. Zwei Punkte sind dabei noch zu betonen:

a) Zusätzlich zu den granularen Kreditdaten, die die OeNB zu den Krediten der Banken erhalten hat und die als Wahrscheinlichkeitsverteilung interpretiert werden können, melden Banken auch noch weitere „aggregierte" Informationen zu ihrem Geschäftsfeld und ihrer Bilanz, z. B. die Bilanzsumme zu einem gegebenen Stichtag. Auch diese Informationen können zum Gruppieren der Banken herangezogen werden. Im Unterschied zu den granularen Kreditdaten liefern diese aggregierten Daten keine allgemeine Wahrscheinlichkeitsverteilung, sondern für jede Bank einen Vektor $x_{agg} \in \mathbb{R}^{d_{agg}}$. Banken B_1, B_2 können gemeinsam anhand der granularen Kreditdaten und der aggregierten Informationen gruppiert werden, wenn als Abstand

$$d(B_1, B_2) := \|x_{1,agg} - x_{2,agg}\|_2 + W_2(\mu_1, \mu_2)$$

genommen wird, wobei $x_{1,agg}$, $x_{2,agg}$ und μ_1,μ_2 jeweils die aggregierten Informationen bzw. die Wahrscheinlichkeitsverteilungen der granularen Kreditdaten der Banken bezeichnen.

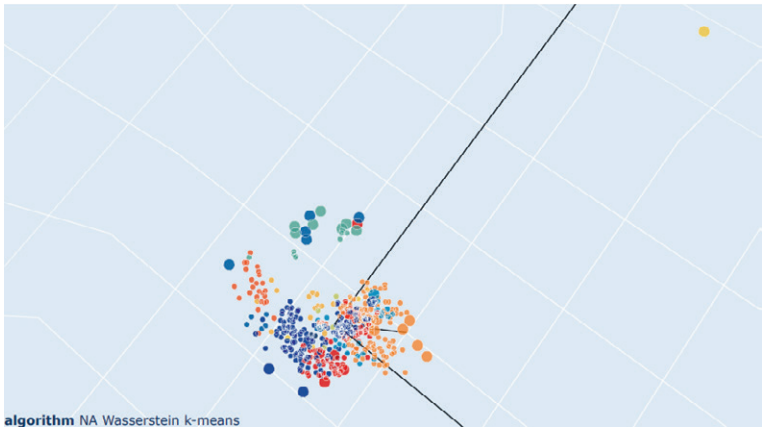
b) Es ist möglich, dass Banken einzelne Kredite – z. B. gegenüber natürlichen Personen – nicht im selben Detailgrad melden müssen wie z. B. Kredite gegenüber Rechtsträgern. Im Extremfall kann es etwa sein, dass ein Attribut der vergebenen Kredite einer Bank nicht bekannt ist, für eine andere Bank jedoch schon. Mathematisch bedeutet dies, dass die Wahrscheinlichkeitsverteilungen solcher zwei Banken nicht im selben Raum leben, sondern die erste eine Wahrscheinlichkeitsverteilung auf einem Teilraum der anderen ist. Hierfür ist es möglich, k-means zu verallgemeinern, um Banken trotzdem gruppieren zu können, siehe Riess et al. (2023). Sind die Banken anschließend gruppiert, können für eine Bank, für die nicht alle Informationen ihrer Daten verfügbar sind, die Informationen von Banken aus dem gleichen Cluster verwendet werden, um die nicht vorhandene Information zu schätzen und zu imputieren.

Wenn man den vorgestellten Clustering-Algorithmus mit granularen Kreditdaten und weiteren aggregierten Daten durchführt, ist zu beachten, dass die Anzahl der betrachteten Datenpunkte je Bank mit Sorgfalt zu wählen ist, da ansonsten Probleme wie der Fluch der Dimensionalität (Konzentrationseffekte in den paarweisen Distanzen zwischen Banken) auftreten können. Weil es sich beim Clustering um ein Verfahren des Unsupervised Learning handelt, sollten die ausgewählten Daten auch zum Analyseziel passen. Sollen beispielsweise Auffälligkeiten bei Immobilienveranlagungen identifiziert werden, ist es zielführend, passende Daten über Immobilien zu verwenden. Je heterogener die Datenquellen sind, desto schwerer fällt es, ein Clustering-Ergebnis zu interpretieren.

Das folgende Kapitel erläutert, wie das Ergebnis eines mithilfe des obigen Algorithmus gefundenen Clusterings visualisiert werden kann, um eine Bankenlandschaft für Analysezwecke zu erzeugen.

Abbildung 6

Visualisierung einer Bankenlandschaft, gefärbt nach Cluster-Zugehörigkeit



Quelle: OeNB.

6 Erzeugung einer Bankenlandschaft

Sind Banken mit Hilfe des im letzten Kapitel beschriebenen Cluster-Algorithmus gruppiert, kann man die Ergebnisse im dreidimensionalen Raum als Bankenlandschaft visualisieren, um die erhaltenen Cluster besser interpretieren zu können, bzw. Auffälligkeiten bildlich zu identifizieren und zu verstehen. Zur Erzeugung einer solchen Bankenlandschaft werden in einem ersten Schritt Wasserstein-Distanzen zwischen allen Banken berechnet. Es existieren unterschiedliche Methoden, um Banken im dreidimensionalen Raum als Punkte zu platzieren, sodass die euklidischen Abstände zwischen den Punkten möglichst den ursprünglichen Wasserstein-Distanzen entsprechen. Da der Raum der Wahrscheinlichkeitsverteilungen unendlich-dimensional ist, muss damit gerechnet werden, dass eine solche Darstellung in drei Dimensionen nicht perfekt ist. Insbesondere ist es möglich, dass für eine Bank der euklidische Abstand im dreidimensionalen Raum zu anderen Banken etwas größer oder kleiner ausfällt als die ursprünglichen Wasserstein-Distanzen. Trotzdem werden Banken mit ähnlichem Kreditportfolio in der Visualisierung nahe beieinander dargestellt, während stärkere Unterschiede im Kreditportfolio auch zu größeren Abständen in der Visualisierung führen werden.

Für die obige Abbildung 6 wurde die Methodik des „Multidimensional Scalings“ (eingeführt von Kruskal, 1964), verwendet, um die Bankpunkte in drei Dimensionen zu generieren. Zusätzlich ist in Abbildung 6 jeder der zehn Cluster unterschiedlich eingefärbt. Wie angemerkt erscheinen Cluster in der Abbildung nicht strikt separiert, was auf die Dimensionsreduktion zurückzuführen ist.

Die datengetriebene Visualisierung der österreichischen Bankenlandschaft in Abbildung 6 erlaubt es, etwaige Banken zu identifizieren, die sich auf Basis der vergebenen Kredite „anders“ verhalten als der Großteil der Banken. Beispielsweise ist die Bank, die zum gelben Punkt in der rechten oberen Ecke der Abbildung gehört, auffällig und kann als Ausreißer betrachtet werden. Es ist davon auszugehen, dass sich die Meldungen der Einzelkreditinformationen jener Bank stark von allen anderen Banken unterscheiden. Diese Information kann als Ausgangspunkt für weitere expert:innenbasierte Analysen genutzt werden, um die Gründe für diese Auffälligkeit herauszufinden: unter Umständen verfügt die Bank über ein vollkommen anderes Geschäftsmodell oder die Meldedaten sind nicht fehlerfrei. Ebenso können unerwartete gemeinsame Gruppierungen von Banken oder Ausreißer innerhalb eines Clusters wertvolle Informationen liefern und Auffälligkeiten darstellen. In Abbildung 6 sind Banken, die beispielsweise „weiter weg“ vom Großteil der Banken liegen, als „größere“ Punkte dargestellt, damit diese Auffälligkeiten leichter erkennbar sind.

7 Conclusio

Granulare Kreditdaten liefern wichtige Informationen zum Kreditgeschäft von Banken. Fasst man die einzelnen Kredite einer Bank mitsamt den gemeldeten Attributen als Wahrscheinlichkeitsverteilung auf, können Banken mit Hilfe der Theorien der Wasserstein-Distanzen und des optimalen Transports miteinander verglichen und gruppiert werden. Unterschiede der Kreditportfolios der Banken können als Distanzen zwischen den Banken dargestellt und mit geeigneten Methoden in Form einer Bankenlandschaft visualisiert werden. Eine solche Bankenlandschaft spiegelt die einzelnen Kreditportfolios wider und erlaubt es, Auffälligkeiten (wie zum Beispiel Ausreißer) besser zu identifizieren. Die in diesem Bericht vorgestellte Methodik ermöglicht es, den Informationsgehalt granularer Kreditdaten vollständig zu verwenden, um die Kreditvergabe einer Bank zu entschlüsseln, ohne dass Informationen in einem aggregierenden Zwischenschritt verloren gehen.

Literaturverzeichnis

Agueh, M. und G. Carlier. 2011. Barycenters in the Wasserstein space. In: SIAM Journal on Mathematical Analysis. 904–924.

Arjovsky, M., S. Chintala, und L. Bottou. 2017. Wasserstein Generative Adversarial Networks. In: Proceedings of the 34th International Conference on Machine Learning. 214–223.

Bachmann, E., M. Hameter, T. Kemetmüller, C. Leitner, P. Reisinger und S. Brachtl. 2021. Progression der Kreditrisikoanalyse durch AnaCredit und die Granulare Kreditdatenerhebung in Österreich. In: Statistiken – Daten & Analysen Q4/21. OeNB. 49–59.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems, volume 26. Curran Associates, Inc.

Hirsch, B., T. Kemetmüller und M. Lingo. 2020. AnaCredit und die Granulare Kreditdatenerhebung (GKE) in Österreich. In: Statistiken – Daten & Analysen Q1/20. OeNB. 20–25.

Kolouri, S., P. Pope, C. Martin und G. Rohde. 2019. Sliced Wasserstein Auto-Encoders. International Conference on Learning Representations.

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. In: Psychometrika, 1–27.

Riess, L., M. Beiglböck, J. Temme, A. Wolf und J. Backhoff . 2023. The geometry of financial institutions – Wasserstein clustering of financial data. arxiv preprint.