

АНАЛИЗ МЕТОДИЧЕСКИХ И АЛГОРИТМИЧЕСКИХ ВОПРОСОВ ИССЛЕДОВАНИЯ И ПРОГНОЗА ПЕРЕХОДОВ ТЕМПЕРАТУРЫ ЧЕРЕЗ НОЛЬ И СВЯЗАННЫХ С НИМИ ГОЛОЛЕДНЫХ ЯВЛЕНИЙ

В.М. Токарев

*ФГБУ «Сибирский региональный научно-исследовательский
гидрометеорологический институт», Новосибирск*

В рамках единого подхода к анализу и прогнозу физически связанных процессов фазовых переходов воды и перехода температуры воздуха (поверхностей) через 0 систематизируются проблемы соответствующих данных наблюдений, их классификации, кластеризации для логической причинно-следственной привязки к прогнозируемым модельным параметрам атмосферы. Одним из ключевых выделяется фактор отсутствия формализации, исследований и наблюдений за таким опасным и сложным явлением антропогенного происхождения, как гололедица. Представлена авторская типизация совместных факторов перехода температуры через 0 и фазовых переходов в виде гололедицы и гололеда (изморозевые явления не рассматриваются).

Для возможности машинного обучения и прогноза предварительно выполнен большой объем по структурной обработке и построению синхронных массивов выходных характеристик двух прогностических моделей (COSMO-Ru_Sib13 и GFS) и подготовленных данных 4-летних наблюдений по станциям Урало-Сибирского региона в соответствии с предложенной типизацией льдообразующих явлений на поверхностях антропогенного происхождения. Описаны алгоритмические этапы вычислительных экспериментов для получения устойчивых статистических решений в виде логических бинарных деревьев.

Ключевые слова: гололедные явления, прогноз, температура, статистические решения, логические бинарные деревья.

ANALYSIS OF METHODOLOGICAL AND ALGORITHMIC ISSUES OF THE RESEARCH AND FORECAST OF ZERO TEMPERATURE TRANSITIONS AND RELATED GLAZE PHENOMENA

V.M. Tokarev

Siberian Regional Research Hydrometeorological Institute, Novosibirsk

Within the framework of a unified approach to the analysis and prediction of physically related processes of phase transitions of water and the transition of air

(surface) temperature through zero, the problems of relevant observational data, their classification, and clustering are systematized for a logical causal link to the predicted model parameters of the atmosphere. One of the key factors is the lack of formalization, research and observation of such a dangerous and complex phenomenon of anthropogenic origin as an ice crusted ground. The author's typification of the joint factors combining temperature transition through zero and phase transitions in the form of ice crusted ground and glaze is presented (rime phenomena are not considered).

For the possibility of machine learning and forecasting, a large amount of structural processing and building of synchronous arrays of output characteristics of two forecasting models (COSMO-Sib and GFS) as well as prepared data of 4-year observations at stations of the Ural-Siberian region was previously performed in accordance with the proposed typification of ice-forming phenomena at surfaces of anthropogenic origin. The algorithmic stages of computational experiments for obtaining stable statistical solutions in the form of logical binary trees are described.

Key words: *glaze phenomena, forecast, temperature, statistical solutions, logical binary trees.*

1. Гололед и гололедица

Гололед – очень опасное метеорологическое явление, поскольку покрывает все поверхности на определенной территории, и защититься от этого невозможно, только частично снизить ущерб при быстром реагировании. К счастью, для Урало-Сибирского региона это весьма редкое явление (за исключением районов с наветренной стороны Уральских гор), требующее тонкого сочетания термодинамических условий в нижних слоях атмосферы.

Гололедица – отдельное следствие определенных погодных условий во всем списке гололедно-изморозевых явлений. Фактически гололедица – явление антропогенного происхождения. В не затронутой человеком природе она незаметна, поскольку вода стекает, впитывается или испаряется быстрее, чем наступает похолодание и замерзание. Редкие лужицы могут оставаться и замерзать только в скальных углублениях или глубоких следах животных. А вот разновидностей гололедицы в городах может быть много, и наблюдается она слишком часто для коммунальных служб, чтобы с нею могли быстро справляться.

Видимо, неопределенность классификации гололедицы из-за множества антропогенных факторов и невозможность их воспроизведения на метеоплощадке не позволили включить ее в список наблюдений и в код КН-01. Поэтому авторам пришлось разработать собственный вариант типизации гололедицы в виде комплексов метеоусловий, способствующих ее появлению и сохранению.

Особенности гололедных явлений (ГЯ):

– неочевидная актуальность ГЯ для Сибири превышает привычную проблемность и морозов, и снегопадов – явлений, к которым приспособились все: население, городские и дорожные службы, транспорт и предприятия;

– с ГЯ все по-другому: объективно сложная для потребителей, журналистов и даже самих синоптиков классификация, статистика и перекрестные причинно-следственные связи явлений;

– самое частое, «рукотворное», но наиболее проблемное для прогнозов и защиты от него явление – гололедица;

– и это явление погоды, приносящее самые массовые травмы населению, аварии и материальные потери автотранспорту, коммунальным и дорожным службам – не наблюдается и не фиксируется в оперативных сводках погоды!

Список наблюдаемых гололедных явлений на метеостанциях разумно ограничен фиксацией явлений природного характера. А потому данных о гололедице нет. Их можно объективизировать посредством кодировки комплекса необходимых условий для ее образования или сохранения.

Авторам пришлось разработать дополнительную типизацию гололеда и морфологическую – для гололедицы, взяв за основу априорные потенциальные метеоусловия и динамику погоды, влияющие на образование, сохранение и ослабление гололедных явлений.

Разработанная первичная типизация гололедицы включает варианты:

1. Замерзание после жидких осадков:

1а) быстрое замерзание, почти все поверхности;

1б) «местами». Замерзание лужиц после стекания, частичного высыхания.

2. Замерзание после оттепели (снежный покров):

2а) быстрое замерзание, почти все поверхности;

2б) «местами». Замерзание лужиц, мокрого снега после стекания, частичного высыхания.

3. «Скрытая». Тонкий свежий снежный покров после типов 1 или 2.

4. «Местами а». Замерзание после дневной «солнечной» оттепели (–3... –7 °C).

5. «Местами б». Стабильно-морозная без осадков погода. Постепенное (до нескольких суток) уплотнение до скольжения неочищенных пешеходных и автодорожных участков.

Варианты 1 и 2 – самые очевидные для осени и весны, а с остальными все сложнее. Неоднозначность «скрытого» варианта 3 следует из характеристик выпавшего снега, его толщины, слипания и пр. Но для пешеходов этот случай – один из самых опасных из-за скрытой непредсказуемости. Также неожиданными, а потому особо опасными могут быть

пятнистые подтаивания на солнце с последующим подмерзанием отдельных участков дорог и тротуаров (вариант 4).

Особую морфологию имеют постепенные уплотнения неубранного снега пешеходами и автотранспортом (вариант 5). Самые травматические и аварийные места хорошо известны (подходы к крупным торговым центрам, остановкам транспорта и, конечно, зоны торможения перед светофорами, перекрестками, спусками. Причем зачастую более опасные места формируются даже не там, где совсем снег не убран, а там, где это сделано запоздало и/или неаккуратно. Усугубляются последствия этого варианта тем, что он может повторяться многократно в течение всего длинного сезона с выпадением снега.

Еще больше запутывает динамику гололеда и гололедицы «жизненный цикл». Многообразие природных факторов (осадки, солнце, ветер, температура, фазовые переходы) и антропогенные «усилия» за и против скольжения ставят почти неразрешимые задачи для прогнозирования опасностей и рисков, связанных с гололедными явлениями. Поэтому авторы надеются не решить, а хотя бы привлечь внимание к проблеме – не «где светлее», а где «ближе к травматологии».

Варианты типизации послужили основой для алгоритмической формализации и детального кодирования погодных ситуаций (табл. 1).

В табл. 2 представлены частотные характеристики расширенных кодов гололедных явлений за 2014–2017 гг. по станциям Урала, Сибири и Якутии.

Максимальные и средние по станциям частоты имеют столь большой разброс, что это требует отдельного методического рассмотрения, причем есть вопросы и к наблюдательной сети (гололед КН-01). Ясен, однако, масштаб влияния явлений поверхностного фазового перехода воды на здоровье и безопасность людей, не говоря уже о транспортных коллапсах и потерях. Просто это почему-то считается неизбежным, как восход-заход солнца. Конечно, на станции Нырб дни совсем без условий для гололедицы найти трудно, но и в урало-сибирских мегаполисах их хватает.

Рассмотрим картирование по разработанным кодам условий погоды и наблюдаемого гололеда (рис. 1–9, табл. 3).

Картирование в целом показывает ожидаемые максимумы гололеда на наветренных западных склонах Уральского хребта, а частоты расчетных кодов трансформации-ослабления существенно меньше. Восточнее Урала частота гололеда резко снижается. Повторяемость условий гололедицы (всех типов) высокая на большей части южной Сибири, Якутии и предгорий Алтая. Заметны и местные (видимо орографические) различия станций.

Большие частотные различия в данных для статистического обучения (распознавания класса предиктанта) требуют специальных алгоритмических подходов. Как правило, а не исключение, редкие явления и погодные условия – самые востребованные объекты прогнозирования и

Таблица 1

Расширенное кодирование гололеда и гололедицы: условия и расшифровка

Код	Гололед
1	Гололед: без условий, берется из сводок КН-01
3	Гололед сохраняется. Условия: после кода 1, пока снег ≤ 1 см, $tn \leq -1$, $td \leq -5$ (до +2 сут.). Расшифровка: ожидается сохранение гололеда до 2 сут без существенного выпадения снега и таяния по температурным порогам дня, ночи (td , tn)
5	Гололед скрытый. Условия: после кодов 1, 3 лег снег ≤ 5 см (до +1 сут.). Расшифровка: ожидается, что под выпавшим до 5 см снегом еще сутки остается опасным сформировавшийся гололед, а далее снег слеживается, начинает утаптываться
8	Гололед слабеет. Условия: после кодов 1, 3, 5 лег снег > 5 см (до +2 сут.). Расшифровка: принимается, что под выпавшим более 5 см снегом гололед еще 2 сут. остается местами опасен для скрытого скольжения
9	Гололед тает. Условия: после кодов 1, 3 $tn > 0$, $td > -5$ (до +1 сут.). Расшифровка: принимается, что за сутки с температурой ночью выше 0°C , а солнечным днем выше -5°C гололед почти везде оттаивает до сравнительно безопасного состояния
Код	Гололедица
2	Гололедица сильная. Условия: жидкие/смешанные осадки + резкий переход через 0°C ($t \leq -10/12$ ч). Расшифровка: принимается, что если после жидких (смешанных) осадков температура за 12 ч опускается ниже -9°C , то формируется сильная (очень опасная) гололедица
4	Гололедица умеренная. Условия: жидкие/смешанные осадки + постепенный переход через 0°C ($t < -7/24$ ч). Расшифровка: принимается, что если после жидких (смешанных) осадков температура за 24 ч опускается ниже -7°C , то формируется умеренная (опасная) гололедица
6	Гололедица местами. Условия: устойчивый снежный покров + оттепель + резкий переход через 0°C ($t \leq -5/12$ ч). Расшифровка: принимается, что после оттепели при устойчивом снежном покрове понижение температуры за 12 ч до -5°C формирует местами опасную гололедицу
7	Гололедица местами – солнце. Условия: устойчивый снежный покров + солнце + $td \geq -5$ + $tn < -5$. Расшифровка: принимается, что после солнечного дня теплее -5°C при устойчивом снежном покрове похолодание ночью ниже -5°C местами формирует опасную гололедицу

Таблица 2

**Частота кодов авторской типизации гололеда и гололедицы
по погодным условиям на территории Урало-Сибирского региона.
Среднегодовые (2014–2017) частоты:
средние по всем станциям и максимумы**

Код – интерпретация	Среднее станций	Максимум станций	Индекс	Метеостанция
1 – гололед (КН-01)	2	44	23912	Ныроб 60.7 56.8 171м Северный Урал
3 – гололед сохраняется	1	24	23912	Ныроб 60.7 56.8 171м Северный Урал
5 – гололед скрытый	0	14	23912	Ныроб 60.7 56.8 171м Северный Урал
8 – гололед слабеет	0	6	23912	Ныроб 60.7 56.8 171м Северный Урал
9 – гололед тает	0	7	23912	Ныроб 60.7 56.8 171м Северный Урал
2 – гололедица сильная	17	157	30961	Оловянная 50.9 115.6 584м Забайкалье
4 – гололедица умеренная	29	175	30961	Оловянная 50.9 115.6 584м Забайкалье
6 – гололедица местами	13	50	23527	Саран-Пауль 64.3 60.9 Ямал
7 – гололедица – солнце	3	15	28144	Верхотурье 58.9 60.8 126м Урал

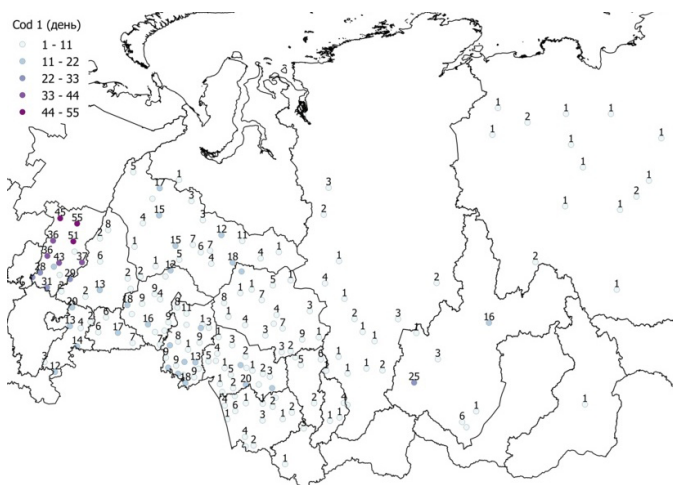


Рис. 1. Распределение частот типов гололеда по метеостанциям за период 2014–2017 гг.: код 1 – дневное время

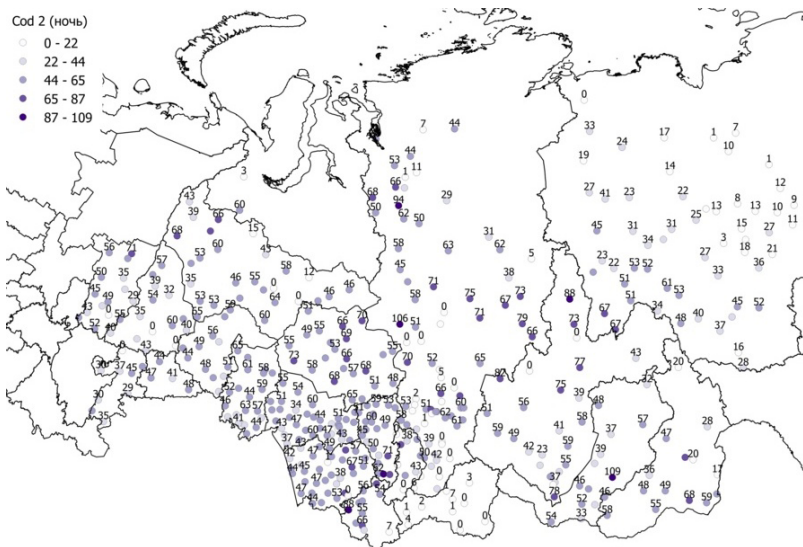


Рис. 2. Распределение частот типов гололедицы по метеостанциям за период 2014–2017 гг.: код 2 – ночное время



Рис. 3. Распределение частот типов гололеда по метеостанциям за период 2014–2017 гг.: код 3 – ночное время

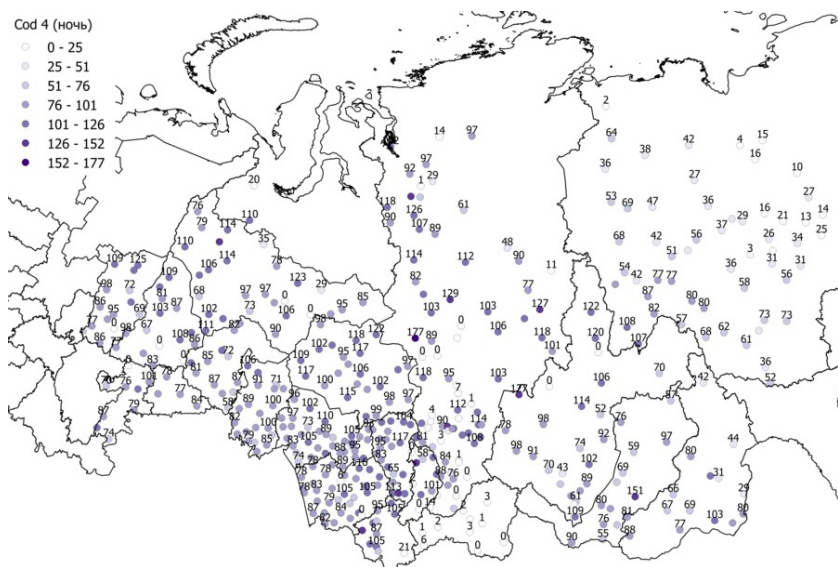


Рис. 4. Распределение частот типов гололедицы по метеостанциям за период 2014–2017 гг.: код 4 – ночное время



Рис. 5. Распределение частот типов гололеда по метеостанциям за период 2014–2017 гг.: код 5 – ночное время

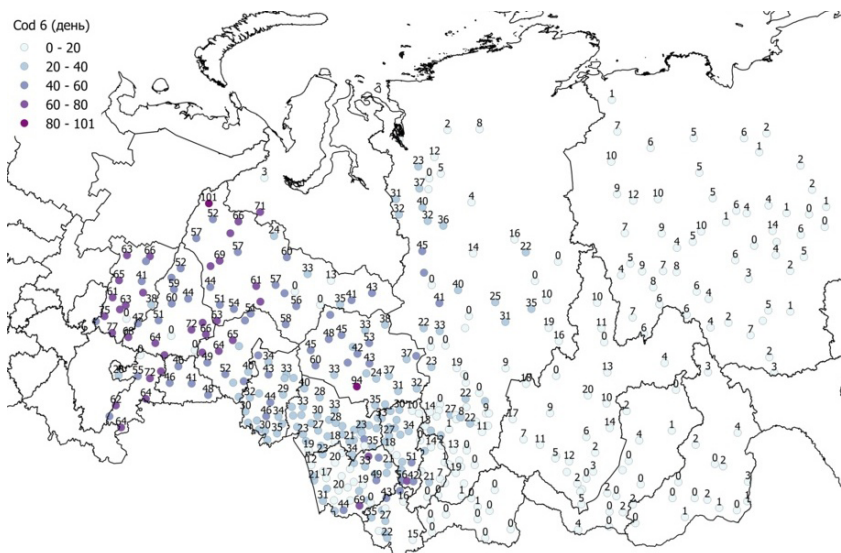


Рис. 6. Распределение частот типов гололедицы по метеостанциям за период 2014–2017 гг.: код 6 – дневное время

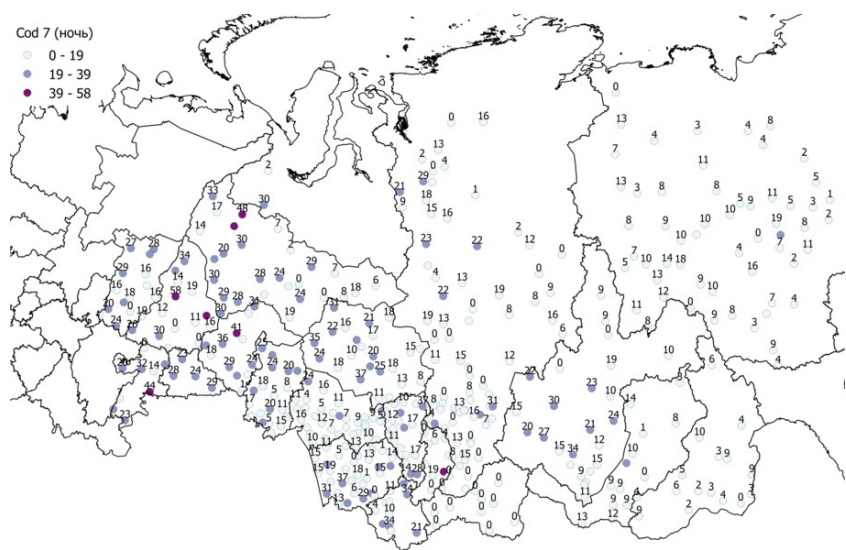


Рис. 7. Распределение частот типов гололедицы по метеостанциям за период 2014–2017 гг.: код 7 – ночное время

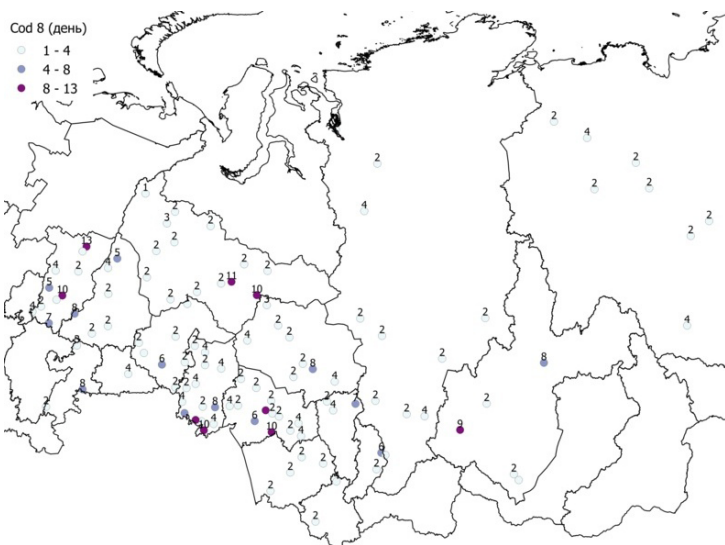


Рис. 8. Распределение частот типов гололеда по метеостанциям за период 2014–2017 гг.: код 8 – дневное время

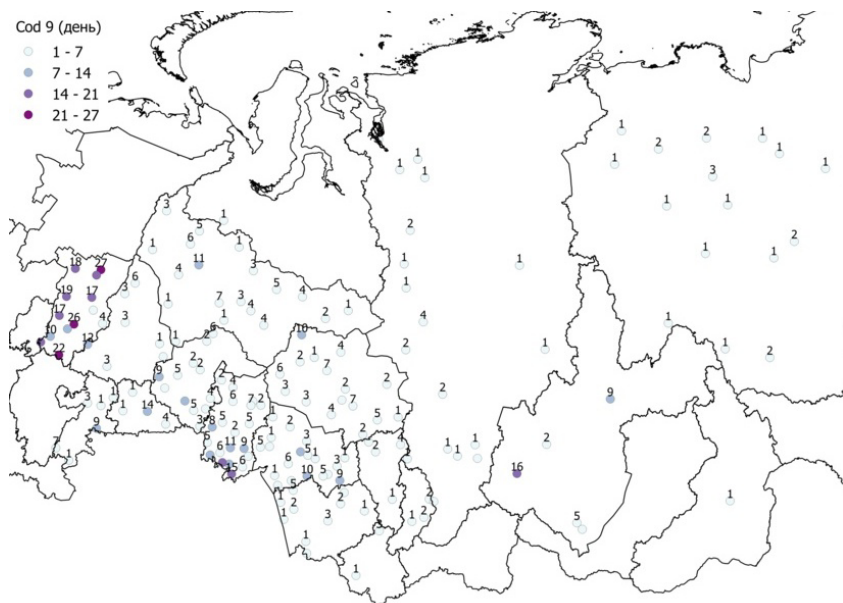


Рис. 9. Распределение частот типов гололеда по метеостанциям за период 2014–2017 гг.: код 9 – дневное время

Таблица 3

Результаты анализа картированных 4-летних частот расширенных кодов гололеда и гололедицы станций Урало-Сибирского региона и Якутии

Код	Гололед	Карта	Анализ 4-летних частот кодов станций на карте
1	Гололед (КН-01)	Рис. 1,	Выше 10 – все Предуралье (день до 55, ночь до 47), редко – Западная Сибирь
3	Гололед сохраняется	Рис. 3,	День 8–24, ночь 20–70 – только Предуралье
5	Гололед скрытый	Рис. 5,	День 4–18, ночь 7–39 – только Предуралье
8	Гололед слабеет	Рис. 8,	Выше 7 (день), 9 (ночь) – только Предуралье, редко – по югу Урала, Сибири
9	Гололед тает	Рис. 9,	Выше 10 (день) – только Предуралье
Код	Гололедица	Карта	Анализ 4-летних частот кодов станций на карте
2	Гололедица сильная	Рис. 2,	Ночь выше 50 – от Урала до юга Якутии, от Горного Алтая, Бурятии до Путорана
4	Гололедица умеренная	Рис. 4,	День выше 50 – юг Сибири, Якутии; ночь выше 80 – Урал, Сибирь, Якутия
6	Гололедица местами	Рис. 6,	Выше 40 (день), 50 (ночь) – Урал, местами
7	Гололедица местами – солнце	Рис. 7,	Западная Сибирь Ночь выше 30 – местами Урал, Западная Сибирь

предупреждения, но основная масса статистических исследований посвящена, напротив, большим, однородным и желательно нормально распределенным характеристикам. К тому же нормативные документы и отсутствие опыта потребителей метеоинформации пока не позволяют использовать вероятностные формулировки прогнозов, которые адекватны природе пространственно-временной изменчивости метеорологических характеристик, условий и явлений погоды. В этом случае детерминированные оценки принято сопровождать формальными дополнениями вида «временами, местами», и каждый потребитель волен трактовать эти скрытые вероятности по своему усмотрению.

Не меньшие сложности это создает и при разработке методов прогнозов, и при их оценке. Очевидно также, что при неравнозначности потенциальных потерь от пропущенного опасного явления и ложного предупреждения методики вынужденно должны подстраиваться в сторону большей предупрежденности и мириться с низкой оправдываемостью прогнозов с явлениями. Для редких ситуаций и явлений асимметрия оценок еще больше, а статистическая обеспеченность прогностических связей еще меньше.

2. О построении логических решающих правил на базе бинарных деревьев

Для бинарного варианта прогноза есть смысл применить процедуру распознавания образов. В нашем случае в качестве базового использован модифицированный вариант алгоритма DW [1], опыт работы с которым имеется в нескольких разработках СибНИГМИ [2, 3].

Алгоритм построен на последовательном делении исходной выборки, содержащей архивные данные двух классов (образов) – с прогнозируемыми явлениями (1-й образ) и без них (2-й образ). Доля случаев с явлениями в исходной выборке принимается за безусловную («климатическую») оценку вероятности 1-го образа. Связь частот (вероятностей) с каким-либо синхронным признаком определяется делимостью условных распределений двух образов. Параметрические методы распознавания используют сравнения статистических оценок (средние, дисперсии). Алгоритм DW непараметрический и ищет условие разделения выборки среди всех значений вариационного ряда признака.

Отметим, что все авторские варианты алгоритма DW используют удобную бинарную форму – «матрицу сопряженности» (табл. 4), из которой в метеорологии получают различные оценки бинарных прогнозов (2.1–2.9). В алгоритме DW для каждого значения признака как порогового (сравнение на « \leq » или « $>$ » порога) для разделения выборки суммируются счетчики клеток матрицы сопряженности в логичном предположении, что полученная подвыборка с большей вероятностью 1-го образа содержит «прогнозы с явлением», а другая – «прогнозы без явления».

Оценки успешности для альтернативных прогнозов согласно табл. 4:

$pr1 = k11/k01$ – предупрежденность наличия явления (2.1),

$pr2 = k22/k02$ – предупрежденность отсутствия явления (2.2),

$vr1 = k11/k10$ – оправдываемость прогнозов наличия явления (2.3),

$vr2 = k11/k10$ – оправдываемость прогнозов отсутствия явления (2.4),

$vr = (k11 + k22)/k00$ – общая оправдываемость прогнозов (2.5),

$LT = k12/k10$ – доля ложных предупреждений (тревог) (2.6),

$TSS = k11/k01 - k12/k02$ – критерий Пирси–Обухова (2.7).

Таблица 4

Таблица сопряженности прогноз – факт

Прогноз	Факт		Сумма
	Да	Нет	
Да	k11	k12	k10
Нет	k21	k22	k20
Сумма	k01	k02	k00

Авторские критерии: PRV и MPR

$PRV = 0.5(pr1 + vr1) - 0.2(pr1 - vr1)$, если $pr1 \geq vr1$,

$PRV = 0.5(pr1 + vr1) - 0.4(vr1 - pr1)$, если $pr1 < vr1$ – критерий баланса (2.8).

PRV-критерий построен с целью сбалансировать две разнонаправленные традиционные характеристики матрицы сопряженности: предупрежденности явления $pr1$ (2.1) и оправдываемости прогнозов наличия явления $vr1$ (2.3).

$vr1$ фактически отражает прогностическую вероятность явления, так мы ее и будем называть. PRV из 2.8 можно было представить в расчетном виде:

$PRV(pr1 \geq vr1) = 0.3pr1 + 0.2vr1$ и $PRV(pr1 < vr1) = 0.9pr1 + 0.6vr1$,

но тогда не виден смысл весовых коэффициентов, а он заключается в балансировке относительно арифметического среднего $pr1$ и $vr1$.

Принимая априори лучшим вариантом равенство предупрежденности и прогностической вероятности ($PRV = pr1 = vr1$), вычитаем отклонение их от равенства с настраиваемым весовым коэффициентом от среднего. Несимметричные веса дают некоторый приоритет (сдвиг) выбранному параметру ($pr1$ или $vr1$).

$MPR = \max \{ \min(pr1, vr1) \}$ – критерий максимизации минимальной предупрежденности (2.9).

Еще один критерий MPR (2.9) был использован также для варианта балансировки большого климатического размаха между данными разных годов. Это максиминный критерий, его максимизация не допустит слишком большого проседания оценок на разных оценочных выборках, правда, за счет некоторого снижения общей (суммарной) оценки.

Осталось свести матрицу сопряженности к какому-то одночисловому критерию качества разделения для простой максимизации. Авторами ранее уже был выбран критерий Пирси–Обухова как приемлемый именно для сравнения разделенных подвыборок, поскольку эксперименты с различными метеорологическими данными показали хорошее взвешенное разнесение вероятностей при максимизации данного критерия.

Итак, на первом шаге алгоритм сначала находит для заданного признака лучшее по критерию Пирси–Обухова пороговое значение для разделения выборки на две с взвешенным разнесением исходной вероятности 1-го образа на большую («прогнозы с явлением») и меньшую («прогнозы без явления»). Полученный результат уже является готовой мини-методикой прогноза по единственному априори заданному признаку. Далее та же процедура повторяется для каждого проверяемого признака с одновременным отбором признака с максимальным значением критерия Пирси–Обухова.

В итоге на первом шаге алгоритма выбран лучший признак из списка с лучшим для него порогом разделения исходной выборки и получены две подвыборки с большей и меньшей вероятностями 1-го образа относительно исходной. Как показал наш опыт при разработке прогнозов гроз, даже такой на вид простой результат может оказаться «лучшим» для очень редких явлений по одному пункту (1–2 случая за 4 года). Понятно, что обеспеченность такой статистики очень низка и нужно менять подход к формированию выборки.

Второй шаг алгоритма повторяет первый два раза – для каждой из двух полученных подвыборок. Последующие шаги выполняются аналогично с удвоением количества обрабатываемых подвыборок.

Таким образом, получаем бинарное дерево цепочек логических выводов вида «если... то... иначе...» с различными вероятностями явления в каждом узле дерева. Деление каждой ветви может заканчиваться по исчерпанию одного из образов, но разумнее предусмотреть критерий останова раньше. Надо хорошо понимать, что с уменьшением объема делимых подвыборок снижается статистическая надежность, устойчивость получаемых результатов вплоть до случайных совпадений.

В принципе, в таком виде «деревом решений» можно пользоваться для выдачи прогнозов в вероятностной форме, поскольку каждый конечный узел имеет вероятность явления, но, чтобы перейти к бинарной формулировке, надо определить вероятностный порог для отнесения прогноза к 1-му или 2-му образу. Один из простых вариантов – сравнивать с начальной «климатической» вероятностью. А далее объединить все матрицы сопряженности конечных парных веток в одну согласно выбранному порогу. При этом вполне могут найтись лишние деления, поскольку они для этого порога дают совпадающие образы, хотя и с различной вероятностью. В любом случае полученная суммарная матрица сопряженности даст оценку качества прогнозов на этой обучающей выборке с заданным вероятностным порогом отнесения любого сочетания используемых признаков к одному из образов.

Кажется очевидным, что оценивать суммарную матрицу сопряженности можно тем же критерием, что и при делении ветвей, но эксперименты с глубокими деревьями показали, что это не лучший вариант.

Существенная авторская модификация алгоритма предусматривает еще 2 шага алгоритма для оптимизации «глубины» дерева, т. е. числа делений каждой ветви и автоматического выбора пороговой вероятности. Для этого вместо произвольного задания глубины и вероятностного порога предложена следующая процедура.

Для каждой глубины дерева, начиная с максимальной, вычисляются матрицы сопряженности для всех вероятностей из конечных ветвей, используя их как пороговые. В результате максимизации критерия находим

как лучшие пороги для каждого варианта глубины дерева, так и лучший вариант его глубины. Вот именно такое детальное сравнение суммарных матриц сопряженности и соответствующих им критериев натолкнуло на необходимость использования критериев, отличных по свойствам от критерия Пирси–Обухова. Первым был вариант с предложенным критерием PRV (баланс предупрежденности и прогностической вероятности явления) (2.8) [2, 3], а позже – критерий минимальной предупрежденности двух образов MPR (2.9). Оба использовались для сравнения различных вариантов построения прогностических решений, выбора лучших для последующей селекции на независимой выборке и оценки оперативных испытаний методик прогнозов.

3. Методология получения комплекса решающих правил и выбора оптимального

Описанный выше модифицированный алгоритм DW на выходе получает субоптимальное (оптимальное потребовало бы перебора всех возможных сочетаний признаков, что неприемлемо) последовательное решение одновременно и для пороговых вероятностей, и для глубины дерева, формализуя субъективные правила останова.

Решение единственно, что для математика – идеальный законченный вариант. А для целей прогноза – это плохо. Важно понимать – почему?

В метеорологии давно принято считать очевидным методологическое правило: после получения статистического (обычно прогностического) решения на некоторой многолетней архивной выборке проверять его на другой, независимой. Это делается для оценки достоверности и устойчивости найденных статистических связей, которые могут быть и вовсе случайными. Оценки на независимой выборке ожидаемо снижаются, и остается лишь оценить их приемлемость для оперативной работы. Но не стоит думать, что математика об этом «не знала».

Изящное обоснование проблем познания дала знаменитая теорема Геделя о неполноте. Из нее, в частности, следует, что выводы, полученные в замкнутой системе, не обязательно справедливы за ее пределами. А ведь ограниченная выборка данных – это и есть наша «замкнутая система». Конечно, и без Геделя было понятно, что оценочные (полученные на ограниченной выборке) статистические связи – это не законы природы, и сила статистики резко слабеет при невыполнении многих оговоренных в ее основах условий: однородности, эргодичности, нормальности и прочего, что обеспечивает строгость ее выводов. Но с Геделем комфортнее.

Можно понять математиков, которые развивают математический аппарат, оперируя модельными данными с заданными свойствами. Так, если выборка данных внутренне неоднородна, то ее надо оценить, описать

математически, отфильтровать и прогнозировать отдельно, как, например, тренд временного ряда. А поскольку в метеорологии мы имеем дело с данными «природы», а не заданного эксперимента, то приходится решать дилемму: или подбирать/разрабатывать адекватный «грязным» данным алгоритм, или подготавливать особым образом данные для применимости выбранного алгоритма. Оба подхода имеют свои преимущества и недостатки, но ничто не запрещает их комплексировать для получения максимального результата.

Примером такого комплексного подхода служат алгоритмические исследования, связанные аббревиатурой МГУА (метод группового учета аргументов) [4]. В обоснование предложенных подходов к обработке данных и прогнозированию положены два принципа: «внешнего дополнения» (явный отклик на теорему Геделя) и «свободы выбора» Габора. Первый принцип относится к формированию данных, а конкретно к делению выборки на части с применением к каждой различных алгоритмов, а второй дает механизм «селекции, самоорганизации», позаимствованный из «дерева жизни». Вот такой любопытный алгоритмический комплекс.

Еще один, более близкий по времени и тематике, пример – моделирование ансамблей в оперативных гидродинамических моделях. Здесь подход несколько иной, но близок в главном – вариации входных данных имитируют различающиеся выборки («внешнее дополнение»), а соответствующее им множество выходных решений позволяет выбрать наиболее вероятное («свобода выбора», селекция).

Аналогичные подходы использованы нами для алгоритмического расширения методики построения логических решающих правил для прогнозирования.

Использовались три выборки многолетнего архива: 2014–2017, 2018, и 2019 гг.

Первая, 4-летняя – для построения базовых деревьев решений глубиной от 1 до 5 уровней для различной полноты списков (9 вариантов) входных параметров. Для последующей селекции по критерию MPR из полученной матрицы (5×9) решений отбирались 9 лучших – матрица (3×3). Вторая (2018 г.) – для селекции лучшего решения по критерию MPR из матрицы решений (3×3). Третья (2019 г.) – для независимой оценки полученных решений.

Алгоритм получения редуцированных списков входных параметров состоял из последовательных шагов построения деревьев, начиная с полного начального списка – 35 прогностических модельных параметров. Для каждого последующего шага список сокращался исключением наименее часто попадавших в деревья параметров с учетом весов, обратно пропорциональных глубинам дерева, на которых признак включался базовым алгоритмом DW. Число единовременных сокращений количества параметров уменьшалось примерно пропорционально текущим размерам спис-

ка от 5–7 на первом шаге до 1 на последнем. Таким образом было получено 9 частотных списков входных параметров.

Алгоритм получения редуцированных деревьев различной глубины (сложности) построен на последовательном обходе узлов выходных деревьев, полученных базовым алгоритмом DW, по уровням глубины, начиная с максимального, и пересчете суммарной матрицы сопряженности для редуцированного на единицу глубины дерева.

Таким образом получено 5 вариантов деревьев решений различной глубины (сложности).

В результате получилась матрица (5×9) вариантов деревьев решений, из которых по критерию MPR отбирались (3×3) лучших и подавались на вход селекции на независимой выборке для получения единственного решения по максимуму MPR.

Описанная иерархическая структура алгоритма направлена на повышение устойчивости решений за счет исключения случайных или редких связей факторов с предиктантом и механизма сбалансированного критерияльного упрощения решений.

Кроме алгоритмических механизмов на различных этапах исследований и вычислительных экспериментов, менялись подходы к формированию входных выборок предикторов и предиктанта. Это определялось как частотными различиями (климат) предиктантов (гололед оказался самым редким из исследуемых явлений), так и надежностью получаемых связей. Поэтому деревья решений строились и по отдельным станциям, и по кластерам с радиусом 200 км, и по территории региона в целом.

Различные подходы к формированию выборки, адекватной особенностям предиктанта, зачастую дают гораздо больше, чем самые навороченные алгоритмические ухищрения. Что и подтверждалось всегда в наших исследованиях.

Литература

1. *Манохин А.Н.* Алгоритм DW для распознавания образов: Пакет прикладных программ ОТЭКС. Новосибирск: Изд-во Новосибирского государственного университета, 1981. С. 3–30.
2. *Здерева М.Я., Токарев В.М.* Анализ и прогноз условий погоды, влияющих на концентрацию атмосферных примесей мегаполиса // Труды СибНИГМИ. 2011. Вып. 106. С. 152–158.
3. *Здерева М.Я., Токарев В.М., Хлущина Н.А., Воробьева Л.П., Бабошина Н.А.* Оперативная технология прогноза гроз в Сибири и результаты ее испытаний // Труды Гидрометеорологического научно-исследовательского центра Российской Федерации. 2018. № 2 (368). С. 27–43.
4. *Ивахненко А.Г., Юрачковский Ю.П.* Моделирование сложных систем по экспериментальным данным. М.: Радио и связь, 1987. 120 с.