Кшнясев И.А. Анализ обилия организмов: мультимодельный вывод как альтернатива проверки нуль-гипотезы // Биологические системы: устойчивость, принципы и механизмы функционирования. Ч.1: сб. материалов III Всеросс. Науч.-практ. Конф. С междунар. участием. Н-Тагил 1-5.03.2010. Нижний Тагил, 2010. С. 348-353.

И. А. Кшнясев
Институт экологии растений и животных УрО РАН
620144, Екатеринбург, ул. 8-го Марта, 202
kia@ipae.uran.ru

АНАЛИЗ ОБИЛИЯ ОРГАНИЗМОВ: МУЛЬТИМОДЕЛЬНЫЙ ВЫВОД КАК АЛЬТЕРНАТИВА ПРОВЕРКЕ НУЛЬ–ГИПОТЕЗЫ

Обычно, в экологических исследованиях анализируются результаты наблюдений (а не управляемых рандомизированных экспериментов) и для интерпретации данных крайне важно иметь «хороший» и даже избыточный, с субъективной точки зрения исследователя, список конкурирующих гипотез (моделей). Итогом статистического анализа является отнюдь не неоспоримая истина, но лишь полезные (оптимальные в некотором смысле) модели исследуемого явления, которые могут быть сопоставлены между собой (на основе поддержки данными), а в дальнейшем уточнены, дополнены или заменены на лучшие. *МультиМодельный вывод* (ММВ) предполагает также оценку параметров (их ошибок и доверительных интервалов) на основе не единственной модели, а их ряда (Anderson, 2008). На примере анализа реальных данных покажем достоинства парадигмы ММВ в ситуации присутствия выборочных нулей и сверхдисперсии дискретных (счетных) данных. Алгоритмы и основная библиография приведены ранее (Кшнясев, 2009), а для экономии объема оценки параметров здесь приведены не будут.

При анализе обилия исследователи имеют дело со счетными данными – числом учтенных особей на некоторой пробной площади. Важным моментом корректного выбора статистического аппарата является спецификация предполагаемого типа распределения зависимой переменной – «обилия» особей. Поскольку признак – дискретный (y_i= 0, 1, 2…), то аппроксимация его нормальным распределением зачастую оказывается не удовлетворительной, а в случае неповторяемых однофакторных планов МНК оценка параметров и их неопределенности (стандартных отклонений) и проверка гипотез оказываются невозможными. Поэтому в таких задачах естественно использовать аппарат теории *Обобщенных Линейных Моделей*, метод *Максимального Правдоподобия* (МП), и соответствующие *дискретные распределения* (Р): Пуассона, Биномиальное, отрицательное Биномиальное и др., в зависимости от типа и способа получения данных (McCullagh, Nelder, 1989). Основанием для выбора может служить оценка согласия статистической модели и данных при спецификации того или иного теоретического распределения. Другими нередкими проблемами являются: *разреженность* данных – наличие выборочных (не структурных) нулей; *эксцесс* – избыток наблюдаемых значений в некоторой категории, по сравнению с ожидаемыми; *инфляция* параметра дисперсии – отклонение оценки от теоретического значения ($\phi \equiv 1$). В статистической

литературе предложен ряд «паллиативных» способов анализа в подобных ситуациях: 1) коллапсирование – объединение некоторых категорий; 2) игнорирование наличия пустых ячеек или использование редуцированных моделей, предполагающих отсутствие взаимодействий; 3) добавление некоторой константы (0.01…1) ко всем (или только к пустым) ячейкам; 4) выбор линк-функций, допускающий нулевые значения отклика ($\eta=\mu^{\lambda}$, λ=0.5); 5) моделирование на основе распределений Пуассона с эксцессом нулевого класса (*ZIP – Zero Inflated Poisson*); 6) отрицательного биномиального; 7) комбинации двух типов распределений: Бернулли и Пуассона (*hurdle model* – «барьерная» модель); 8) применение обычного МНК, после стабилизирующего дисперсию преобразования (z=log(y+c), $z=y^{0,25}$ и др.).

Статистическая линейная модель для счетных данных (при лог-преобразовании) может быть записана:

$$Log[E(y)] = E[\eta] = log(t) + \boldsymbol{\theta'X} \ ,$$

где $E(y_i)$ – ожидаемая число особей, $\boldsymbol{\theta}$ – вектор оцениваемых (неизвестных) регрессионных коэффициентов, $\boldsymbol{X}$ – вектор наблюдаемых значений предикторов, а *log(t)* – переменная (*offset*), коэффициент при которой предполагается известным и равным единице, необходимая в случае различных: исследованной площади, числа ловушек или продолжительности наблюдений. Для контроля согласия модели и данных используют отношение критерия согласия Пирсона (реже, девианс $D=-2logL$) к числу степеней свободы остатка: $\phi=X^2/rdf$. В случае отклонения от теоретической ($\phi\equiv1$) оценки параметра дисперсии, как для проверки H_0–гипотезы, так и для построения доверительных интервалов (Nelder, Wedderburn, 1972) используется модификация традиционных критериев: статистики типа *Вальда*:

$$`X^2\text{-}Вальда = \theta_i^2/\{\phi[SE(\theta_i)]^2\} = \theta_i^2/\phi Var(\theta_i) \ ,$$

где $Var(\theta_i) = [SE(\theta_i)]^2$ – обычная МП оценка параметра дисперсии, ϕ – параметр инфляции дисперсии ($\phi=X^2/rdf$) или типа *Отношения Правдоподобия*:

$$`G^2 = G^2/\phi = [2LL(\theta_i) - 2LL(\theta_0)]/\phi$$

Полученные значения сравниваются с критическими $\boldsymbol{\chi}^2(k_i\text{-}k_0)$, где k_i – число параметров исследуемой модели, а k_0 – число параметров редуцированной (обычно, H_0, содержащей только константу) модели. Доверительные интервалы для параметров также должны учитывать инфляцию дисперсии, например 95% ДИ: $\theta_i \pm 1.96\phi^{0.5}SE(\theta_i)$.

В таблице 1 представлены (y_i) результаты учетов численности (Мухачева и др., 2009) мелких млекопитающих (ММ) методом ловушко-линий на 10 участках, в трех условных зонах техногенной нагрузки, на различном расстоянии в северном и южном направлениях от КМК – Карабашского медеплавильного комбината (Челябинская обл.). Проблема разреженности данных имеет место – 8 из 30 ячеек (26,7%) – выборочные нули – крайне низкая вероятность обнаружения животных. Ясно, что использование логарифмической (канонической для счетных данных) линк-функции невозможно, но возможно степенное (в том числе и λ=1) преобразование. Одна из диагностик – среднее число наблюдений на ячейку, равное 4.3, свидетельствует о том,

что при анализе даже не коллапсированных данных получим относительно удовлетворительную аппроксимацию используемых статистик (G^2, X^2) распределением хи-квадрат. Поскольку данные получены с помощью ловушек-давилок, результаты учетов представляют собой число пойманных животных («успехов») при конкретном числе попыток (ловушко-суток), что явно подсказывает использовать схему Бернулли. Однако, для возможного приложения схемы анализа к данным, получаемым другими способами, например, ловчими цилиндрами («накопительными ловушками») или маршрутным учетом и др. полезно рассмотреть предположение, что результаты учетов имеют распределение Пуассона (или его обобщение). Во всех случаях при спецификации (Р) распределения Пуассона, модель будет учитывать не равное число участков в разных условных зонах загрязнения. Исследуем (табл. 3) и прокомментируем результаты (табл. 4) некоторых способов анализа.

Таблица 1

Число ММ (y_i), отловленных за 3 ночи (3 линии по 25 ловушек) на 10 участках, в трех условных зонах техногенной нагрузки, на различном расстоянии (км) в северном (-) или южном направлениях от КМК (Челябинская обл.); октябрь, 2008 г

Зона техногенной нагрузки (загрязнения)					
Импактная		Буферная		Фоновая	
Км.	y_i	Км.	y_i	Км.	y_i
-1	***0, 0, 0***	9	*4,* ***0,*** *2*	25	*1, 3, 11**
3	***0,*** *2, 1*	-10	*6, 4, 19**	30	*8, 2, 12**
-5	***0, 0,*** *1**	12	*4, 6, 10**	-32	***0,*** *3, 4*
		-18	*8, 5, 13**		

*Примечание: * - бинарный предиктор (0; 1) характеризует локальную гетерогенность, но не будет включен в исследуемые модели, поскольку нас интересует вывод в условиях сверхдисперсии, а так же и из дидактических соображений.*

Таблица 2

Оценка согласия (GOF) моделей и данных, проверка H_0 гипотез об эффектах: «Направление» (А) и «Зона» (В) на обилие ММ в окрестностях КМК, октябрь, 2008 г

Модель и модификация данных	Р	η	**GOF** (не)согласие			$'X^2(df)$ - Критерий Вальда для эффекта		
			rdf	X^2	ϕ	А (*df=1)*	В (*df=2*)	А×В (*df=2)*
1. y>0	КП	log	16	44.37	2.77	0.04	3.79	2.17
2. Не модиф.	КП	λ=0.25	24	61.23	2.55	0.88	**17.33**	**7.60**
3. Коллапс в 10 ячеек	КП	λ=0.25	4	10.03	2.51	0.90	**17.64**	**7.74**
4. Коллапс в 6 ячеек	П	log	0	0	(1)	2.60	**24.36**	**16.31**
5. y+0.5	КП	log	24	52.40	2.18	0.65	**13.59**	**6.72**
6. y+1	КП	log	24	47.87	1.99	0.45	**14.38**	**6.75**
7. если y=0, то y+0.01	КП	log	24	60.82	2.53	1.01	**9.74**	**6.42**
8. y+0.01	Б	logit	2244	2250.6	1.003	2.52	**26.09**	**17.55**
9. Не модиф.	Б	logit	2244	2250	1.003	2.55	**25.68**	**17.57**

Примечание: Р – распределение: П – Пуассон, КП – квази-Пуассон,
Б-Биномиальное. Жирный шрифт – «значимые» эффекты.

Таблица 3

Комментарии к оценке согласия и проверки гипотез см. табл. 2

№	Комментарий
M1	При игнорировании пустых ячеек, 8 линий и один исследованный участок выпадает из анализа, среднее число наблюдений на ячейку плана 5.86; максимальная сверхдисперсия. Самая неудачная модель, статистики (и достигнутые уровни значимости) для эффектов явно не реалистичны.
M2	Использование степенной линк-функции $y^{0,25}$ позволяет включить в анализ и пустые ячейки, среднее число наблюдений на ячейку 4.3, сверхдисперсия.
M3	Коллапсирование 3 линий внутри каждой из 10 площадок, всего одна пустая ячейка, среднее число наблюдений на ячейку 12.9; сверхдисперсия. Изменчивость между линиями внутри участка моделируется оценкой параметра дисперсии.
M4	Объединение линий внутри зон каждого направления, приводит к плану без повторностей но и без пустых ячеек, (т.н. «насыщенная» модель – число параметров равно числу ячеек плана, предсказываемые значения равны наблюдаемым: $X^2=0$, $rdf=0$), среднее число наблюдений на ячейку 21.5 максимально. Игнорируется изменчивость как между линиями внутри участка, так и между участками внутри зоны.
M5, M6, M7	Добавление констант ко всем ячейкам модели типа M2 позволяет использовать лог-преобразование. Результаты M5–M7 качественно сходны, оптимальное значение дисперсии при +1.
M8, M9	Биномиальная модель (с фиксированными эффектами) автоматически объединяет результаты учетов внутри ячеек плана – линий и участков внутри зон одного направления (нет пустых ячеек) насыщенная модель. Игнорируется изменчивость между линиями внутри участка и между участками внутри зоны.

Если придерживаться парадигмы проверки гипотез (H_0), то, как можно заключить из результатов (табл. 2) и комментариев к ним (табл. 3), все варианты, исключая первый, приводят к качественно сходному выводу о важности эффектов «зоны техногенной нагрузки» (удаленности от источника загрязнения) и взаимодействия «направление-зона». К аналогичным результатам приводит и использование информационных критериев, причем, даже используя различную спецификацию типа распределения объясняемой переменной – Биномиальное (табл. 4, M9 из табл. 2) или квази-Пуассона (табл. 5, M1 из табл. 2) – оптимальными признаются одни и те же модели, что свидетельствует о мощности и робастности вывода. Использование же классической парадигмы проверки H_0 (при корректировке на сверхдисперсию) и удалении выборчных нулей (M1, табл. 2) приводит к катастрофической потере *мощности* статистического вывода, информационные же статистики и в таких «плохих» условиях работают адекватно (табл. 2). Оценки «важности» (w^+) для объяснения изменчивости зависимой переменой приведены в примечании к таблицам 4 и 5, и качественно сходны.

Статистические методы, основанные на концепциях теории информации, представляют собой унифицированный инструмент анализа данных, имеют фундаментальную теоретическую основу, позволяют сравнивать неиерархические модели и решать «проблему псевдоповторностей» (табл. 4); подобно байесовским методам оценивать апостериорные вероятности (только при

равных априорных) w – «веса» гипотез. В рассмотренном примере, ММВ характеризуется уникальным сочетанием, как мощности (чувствительности), так и робастности (устойчивости) вывода, причем даже в случае, когда методы, основанные на парадигме проверки гипотез, оказываются неэффективными.

Таблица 4

Отбор оптимальных (*MinCAIC*) логит-моделей изменчивости обилия ММ в окрестностях КМК октябрь, 2008 г. *P – Биномиальное. MinCAIC =943.7*

Предикторы	*K*	*-2LL*	Δ_{CAIC}	*w*
Зона + Зона×Направление	**5**	**900.11**	**0.00**	**0.638**
Зона	**3**	**918.93**	**1.38**	**0.319**
Зона + Направление + Зона×Направление	6	897.12	5.73	0.036
Зона + Направление	4	917.99	9.16	0.007
Участок ϕ=0,9	*10*	*885.67*	*29.15*	*3E-07*
Зона×Направление	3	956.63	39.08	2E-09
Направление + Зона×Направление	4	955.4	46.57	5E-11
H_0	*1*	988.05	53.07	2E-12
Направление	2	987.97	61.70	3E-14
Линия ϕ=0,74	*30*	*826.17*	*144.03*	*3E-32*

Примечание: w^+(Зона)=1, w^+(Зона×Направление)=0.674, w^+(Направление)=0.043.

Таблица 5

Отбор (*MinCAIC*) моделей обилия ММ в окрестностях КМК, октябрь, 2008 г. *P – Квази-Пуассон*, усеченные данные (М1, в табл. 2): y_i>0. *MinCAIC* =138.5

Предикторы	*K*	*-2LL*	*dCAIC*	*w*
Зона	**3**	**126.3**	**0.0**	**0.417**
Зона + Зона×Направление	**5**	**118.1**	**0.0**	**0.413**
Зона + Направление	4	124.8	2.7	0.110
Зона + Направление + Зона×Направление	6	118.0	4.0	0.056
Зона×Направление	3	136.7	10.5	0.002
Направление	2	142.4	12.0	0.001
Направление + Зона×Направление	4	135.1	12.9	0.001
H_0	1	272.5	138.1	4E-31

Примечание: w^+(Зона)=0.996, w^+(Зона×Направление)=0.472, w^+(Направление)=0.169.

СПИСОК ЛИТЕРАТУРЫ

Кшнясев И.А. Информационные критерии и их приложения в анализе экологических данных // Ученые записки НТГСПА 2009. Естественные науки 2008-2009. С. 157-166.

Мухачева С.В., Кшнясев И.А., Давыдова Ю.А. Плотность и структура населения мелких млекопитающих в окрестностях Карабашкого медеплавильного комбината // Современные проблемы зоо- и филогеографии млекопитающих. Матер. Конф. М.: КМК. 2009. С. 57.